Spring 5-31-2013

Novel color and local image descriptors for content-based image search

Sugata Banerji
New Jersey Institute of Technology

Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

NOVEL COLOR AND LOCAL IMAGE DESCRIPTORS FOR CONTENT-BASED IMAGE SEARCH

by
Sugata Banerji

Content-based image classification, search and retrieval is a rapidly-expanding research area. With the advent of inexpensive digital cameras, cheap data storage, fast computing speeds and ever-increasing data transfer rates, millions of images are stored and shared over the Internet every day. This necessitates the development of systems that can classify these images into various categories without human intervention and on being presented a query image, can identify its contents in order to retrieve similar images.

Towards that end, this dissertation focuses on investigating novel image descriptors based on texture, shape, color, and local information for advancing content-based image search. Specifically, first, a new color multi-mask Local Binary Patterns (mLBP) descriptor is presented to improve upon the traditional Local Binary Patterns (LBP) texture descriptor for better image classification performance. Second, the mLBP descriptors from different color spaces are fused to form the Color LBP Fusion (CLF) and Color Grayscale LBP Fusion (CGLF) descriptors that further improve image classification performance. Third, a new HaarHOG descriptor, which integrates the Haar wavelet transform and the Histograms of Oriented Gradients (HOG), is presented for extracting both shape and local information for image classification. Next, a novel three Dimensional Local Binary Patterns (3D-LBP) descriptor is proposed for color images by encoding both color and texture information for image search. Furthermore, the novel 3DLH and 3DLH-fusion descriptors are proposed, which combine the HaarHOG and the 3D-LBP descriptors by means of Principal Component Analysis (PCA) and are able to improve upon the individual HaarHOG and 3D-LBP descriptors for image search. Subsequently, the innovative H-descriptor, and the H-fusion descriptor are presented that improve upon the 3DLH descriptor. Finally, the innovative

Bag of Words-LBP (BoWL) descriptor is introduced that combines the idea of LBP with a bag-of-words representation to further improve image classification performance.

To assess the feasibility of the proposed new image descriptors, two classification frameworks are used. In one, the PCA and the Enhanced Fisher Model (EFM) are applied for feature extraction and the nearest neighbor classification rule for classification. In the other, a Support Vector Machine (SVM) is used for classification. The classification performance is tested on several widely used and publicly available image datasets. The experimental results show that the proposed new image descriptors achieve an image classification performance better than or comparable to other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), the Spatial Envelope (SE), the Color SIFT four Concentric Circles (C4CC), the Object Bank (OB), the Hierarchical Matching Pursuit (HMP), the Kernel Spatial Pyramid Matching (KSPM), the SIFT Sparse-coded Spatial Pyramid Matching (ScSPM), the Kernel Codebook (KC) and the LBP.

NOVEL COLOR AND LOCAL IMAGE DESCRIPTORS FOR CONTENT-BASED IMAGE SEARCH

by
Sugata Banerji

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Computer Science

Department of Computer Science

May 2013

APPROVAL PAGE

NOVEL COLOR AND LOCAL IMAGE DESCRIPTORS FOR CONTENT-BASED IMAGE SEARCH

Sugata Banerji

Dr. Chengjun Liu, Dissertation Advisor — Date
Associate Professor of Computer Science, NJIT

Dr. Durgamadhab Misra, Committee Member — Date
Professor of Electrical and Computer Engineering, NJIT

Dr. David Nassimi, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Dimitrios Theodoratos, Committee Member — Date
Associate Professor of Computer Science, NJIT

Dr. Zhi Wei, Committee Member — Date
Assistant Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Sugata Banerji

Degree: Doctor of Philosophy

Date: May 2013

Undergraduate and Graduate Education:

- Doctor of Philosophy in Computer Science,
  New Jersey Institute of Technology, Newark, New Jersey, 2013
- Bachelor of Engineering in Information Technology,
  West Bengal University of Technology, Kolkata, India, 2005

Major: Computer Science

Publications:

S. Banerji, A. Sinha, and C. Liu, "HaarHOG: Improving the HOG Descriptor for Image Classification," *Pattern Recognition Letters*, (under review).

S. Banerji, A. Sinha, and C. Liu, "New Image Descriptors Based on Color, Texture, Shape, and Wavelets for Object and Scene Image Classification," *Neurocomputing*, in press, 2013.

A. Sinha, S. Banerji, and C. Liu, "New Color GPHOG Descriptors for Object and Scene Image Classification," *Machine Vision and Applications*, (under review).

S. Banerji, A. Verma, and C. Liu, "LBP and Color Descriptors for Image Classification" in *Cross Disciplinary Biometric Systems*, C. Liu and V. Mago Eds., Springer-Verlag, pp. 205-225, 2012.

A. Sinha, S. Banerji, and C. Liu, "Novel Color Gabor-LBP-PHOG (GLP) Descriptors for Object and Scene Image Classification," *The Eighth Indian Conference on Vision, Graphics and Image Processing*, December 16-19, 2012, Mumbai, India.

A. Sinha, S. Banerji, and C. Liu, "Gabor-Based Novel Local, Shape and Color Features for Image Classification," *The 19th International Conference on Neural Information Processing*, November 12-15, 2012, Doha, Qatar.

A. Sinha, S. Banerji, and C. Liu, "Novel Gabor-PHOG Features for Object and Scene Image Classification," in *Structural, Syntactic, and Statistical Pattern Recognition*, Lecture Notes in Computer Science Volume 7626, pp 584-592, 2012.

S. Banerji, A. Sinha, and C. Liu, "Novel Color HWML Descriptors for Scene and Object Image Classification," *The 3rd International Conference on Image Processing Theory, Tools and Applications*, October 15-18, 2012, Istanbul, Turkey.

S. Banerji, A. Sinha, and C. Liu, "Scene Image Classification: Some Novel Descriptors," *IEEE International Conference on Systems, Man, and Cybernetics*, October 14-17, 2012, Seoul, Korea.

S. Banerji, A. Sinha, and C. Liu, "Novel Color, Shape and Texture-based Scene Image Descriptors," *2012 IEEE International Conference on Intelligent Computer Communication and Processing*, August 30- September 1, 2012, Cluj-Napoca, Romania.

S. Banerji, A. Sinha, and C. Liu, "Object and Scene Image Classification Using Unconventional Color Descriptors," *The 16th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, July 16-19, 2012, Las Vegas, Nevada, USA.

A. Sinha, S. Banerji, and C. Liu, "Gabor-Based Novel Color Descriptors for Object and Scene Image Classification," *The 16th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, July 16-19, 2012, Las Vegas, Nevada, USA.

S. Banerji, A. Verma, and C. Liu, "Novel Color LBP Descriptors for Scene and Image Texture Classification," *The 15th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, July 18-21, 2011, Las Vegas, Nevada, USA.

A. Verma, S. Banerji, and C. Liu, "A New Color SIFT Descriptor and Methods for Image Category Classification," *The 2010 International Congress on Computer Applications and Computational Science*, December 4-6, 2010, Singapore.

To My Beloved Family

Who Started and Nurtured

This Lifelong Fascination with Images

ACKNOWLEDGMENT

Firstly, I would like to take this opportunity to express my heartfelt appreciation to my dissertation advisor, Dr. Chengjun Liu, for his invaluable advice, infinite patience, technical guidance, and his confidence in me which enabled me to bring this dissertation to its culmination. Over the past few years, Dr. Liu has been a constant source of encouragement and a valued mentor to me. I will always be indebted to Dr. Liu for encouraging and supporting me to present our research achievements to international conferences and reputed journals, which significantly improved my confidence and gave me great opportunities to exchange ideas with other researchers in my field.

Secondly, I am extremely grateful to Dr. Durgamadhab Misra, Dr. David Nassimi, Dr. Dimitrios Theodoratos and Dr. Zhi Wei for serving on my committee. In addition, I would like to extend special thanks to Dr. Sushmita Mitra from Indian Statistical Institute, Kolkata, India. They have provided me with academic advice, both related to my dissertation topic and otherwise. This dissertation would not have been possible without their invaluable guidance and the time they spent with me. I would also like to thank my fellow graduate students for their assistance and support, in particular Atreyee Sinha, Abhishek Verma, Shuo Chen, and Zhiming Liu. Their friendship and feedback have been a great source of emotional support. Special mention should be made of the fact that Atreyee spent many days proof-reading this dissertation, and helped me draw several of the figures included here.

Most importantly, I heartily appreciate my family for standing by my side not only through the tough years of Ph.D. study, but my entire life. I thank my parents, Mr. Gautam Banerji and Mrs. Paramita Banerji, for their faith in me and allowing me to be as ambitious as I wanted. It was under their watchful eye that I gained so much strength and ability to tackle challenges head-on. Along the way, I am thankful to my sister, Sujata Chatterjee, for providing endless encouragement and helping me retain the child in me.

Finally, I would like to thank the people who have stayed with me for the past five years, giving me advice on non-research related matters and keeping me sane. In particular, I would like to thank my friends: Ankur Agrawal, Kashif Qazi, Amrita Banerjee, Sumana Pai, Dr. Fadi Karaa, Mrs. Shoreh Karaa, and my cousins: Mr. Soumyo Chakraborty, Ms. Pomeli Ghosh, Dr. Tathagata Mukherjee and Ms. Paulami Chatterjee. Last, but not the least, I would like to thank Dr. Jorge Cham, the creator of PhD Comics, for making me realize that I am not alone.

TABLE OF CONTENTS

TABLE OF CONTENTS

(Continued)

TABLE OF CONTENTS

(Continued)

Chapter **Page**

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

The area of content-based image classification, search and retrieval is not new. Image recognition has a wide variety of uses, including but not limited to medical diagnostics, weather prediction, agriculture, security and surveillance, military applications and robot vision. The research on this field has expanded greatly in recent years. With the easy availability of inexpensive digital cameras, cheap data storage, fast processing speeds, ever-increasing data transfer rates, and the growing popularity of social networking and media sharing websites, millions of color images are stored and shared over the Internet each day. According to a report published by the website visualnews.com and summarized in Figure 1.1, the social networking site Facebook currently holds over 140 billion digital

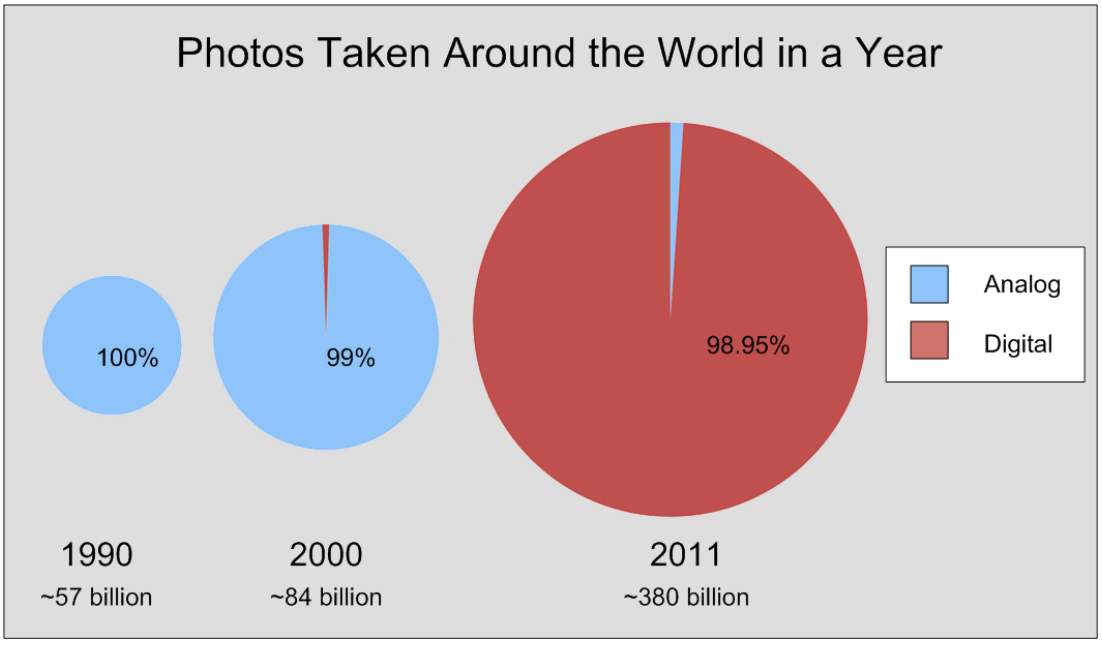


Figure 1.1 The number of photos taken around the world in 1990, 2000 and 2011. Data taken from: P. Caridad, "Smile for the Cell Phone!– New Photography Trends", http://www.visualnews.com/2012/06/11/smile-for-the-cell-phone-new-photography-trends, 2012.

images, to which an average of over 300 million were added every day in March 2012. The photo sharing website Instagram, where 60 photos are added every second, had four billion photos uploaded as of July 2012. An estimated 380 billion images were taken worldwide in 2011 alone, 99% of which were digital. Figure 1.1 shows this sudden growth in the number of digital photos acquired around the world. This large volume of digital images necessitates the development of systems that can classify these images into different categories without human intervention. Also, there is a growing demand for an image-based search system that, on being presented a query image, can identify its contents and retrieve images containing similar elements. Creation of the feature descriptor is one of the first steps in the image search and classification process and this dissertation introduces different image descriptors for color images.

The color cue is often applied by the human visual system for object and scene image classification. Indeed, color images, which contain more discriminative information than grayscale images, have been shown to perform better than grayscale images for image classification tasks (Liu and Mago 2012; Banerji et al. 2011; Liu 2011; Liu and Yang 2009; Liu 2007, 2004). Image descriptors defined in different color spaces usually help improve the identification of object, scene and texture image categories (Verma et al. 2010; Banerji et al. 2011). The descriptors derived from different color spaces often exhibit different properties, among which are high discriminative power and relative stability over the changes in photographic conditions such as varying illumination. Color histogram and global color features and local invariant features often provide varying degrees of success against image variations such as rotation, viewpoint and lighting changes, clutter and occlusions (Burghouts and Geusebroek 2009; Stokman and Gevers 2007).

Texture, shape, and local information contribute as well to object and scene image classification. Local Binary Patterns (LBP), for example, has been shown to be promising for recognition and classification of texture images (Ojala et al. 1994; Zhu et al. 2010; Crosier and Griffin 2008). The Histograms of Oriented Gradients (HOG) descriptor (Dalal

and Triggs 2005), which represents an image by histograms of the slopes of the object edges in an image, stores information about the shapes contained in the image. As a result, HOG has become a popular descriptor for content based image retrieval. In addition, wavelets, such as the Haar wavelets have been widely applied for object detection in images (Zhang et al. 2007c).

This dissertation explores several novel image descriptors based on texture, shape, color and local features from an image. Specifically, first, a new color multi-mask Local Binary Patterns (mLBP) descriptor is introduced that represents texture and achieves better classification performance than traditional LBP descriptors. Second, the mLBP descriptors from different color spaces are fused to form the Color LBP Fusion (CLF) and Color Grayscale LBP Fusion (CGLF) descriptors that perform better than the descriptors from individual color spaces. Third, a new HaarHOG feature vector is introduced that extracts shape as well as local features from an image by combining the Haar wavelet transform with the Histograms of Oriented Gradients (HOG). Fourth, the concept of LBP is extended to generate the novel Three Dimensional Local Binary Patterns (3D-LBP) descriptor for color images that encodes color information along with the texture information from an image. Next, a Principal Component Analysis (PCA) based fusion technique is applied to combine the 3D-LBP and HaarHOG descriptors and the 3DLH and 3DLH-fusion descriptors are generated that outperform both 3D-LBP and HaarHOG descriptors in classifying different types of images. Further, the innovative H-descriptor and the H-fusion descriptor are proposed, the latter of which outperforms the 3DLH-fusion. Finally, a new Bag of Words LBP (BoWL) image descriptor is introduced which not only outperforms the traditional LBP by a big margin, but also performs quite well on scene images.

The classification performance of the proposed descriptors are assessed using two methods. In one, the proposed new image descriptors are subjected to dimensionality reduction by PCA and feature extraction using Enhanced Fisher Model (EFM). Then a nearest neighbor classifier is used to test their performance on several widely used and publicly

available image datasets. In the other method, a Support Vector Machine (SVM) classifier is used for classification. In both types of experiments, the proposed descriptors are shown to achieve a better classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope (SE), Color SIFT four Concentric Circles (C4CC), Object Bank (OB), the Hierarchical Matching Pursuit (HMP), the Kernel Spatial Pyramid Matching (KSPM), the SIFT Sparse-coded Spatial Pyramid Matching (ScSPM), the Kernel Codebook (KC)as well as LBP and a few others.

This dissertation is organized in the following manner. Chapter 2 discusses the related work by other researchers that have been used in this dissertation. Chapter 3 discusses texture and scene image classification by a new mLBP descriptor. Chapter 4 introduces the novel HaarHOG descriptor and evaluates its classification performance. Chapter 5 explains three new descriptors: the 3D-LBP, the 3DLH and the 3DLH-fusion. Chapter 6 introduces two novel descriptors, the H-descriptor and the H-fusion descriptor, that incorporate color, shape, texture and local information from an image. Chapter 7 introduces the BoWL descriptor that uses the LBP concept from Chapter 3 but significantly improves classification performance. Chapters 3, 4, 5, 6 and 7 also include the results of experiments done on various image datasets. A more detailed discussion of the experimental results has been included at the end of Chapter 6 to further evaluate the performance of the H-fusion descriptor on different categories of various image datasets. Finally, Chapter 8 summarizes the contributions of this dissertation and discusses future directions for research.

CHAPTER 2

BACKGROUND

An image is stored digitally as a matrix of values. It can be considered a two-dimensional function $f(x,y)$ defined over the spatial domain where the value of the function at some particular x and y gives the image intensity at that point. Each of these discrete intensity values, i.e. each of these elements of the matrix, is known as a picture element, or "pixel" in short. Color images contain three such intensity matrices and can reproduce colors by storing three intensity values for each pixel of an image.

Although color images contain more discriminative information than grayscale images, the use of full color features as a means to image retrieval (Liu and Mago 2012; Liu 2011; Liu and Yang 2009; Liu 2006) and object, texture and scene search (Verma et al. 2010; Banerji et al. 2011) had not gained popularity until recently. This is because using the complete color information for feature extraction requires high computing power as well as more memory since color images contain at least three times the information contained in grayscale images. Discriminative information can be captured from color images by means of color features such as color invariants, color histogram and color texture. The early methods for object and scene classification were mainly based on the global descriptors such as the color and texture histograms (Niblack et al. 1993; Pontil and Verri 1998; Schiele and Crowley 2000). One such representative method is the color indexing system designed by Swain and Ballard, which used the color histogram for image retrieval from a large image database (Swain and Ballard 1991). These early methods were sensitive to viewpoint and lighting changes, clutter and occlusions. For this reason, alongside global methods, part-based methods were also developed, which became the popular techniques in the object recognition community. Part-based models combine appearance descriptors from local features along with their spatial relationship (Fergus et al. 2003; Fischler and Elschlager 1973). However, learning and inference for spatial relations poses a challenging

problem in terms of its complexity and computational cost.

More recently, the work of (Verma et al. 2010; Liu and Mago 2012), and (Liu 2008) on color based image classification propose several new color spaces and methods for face, object and scene classification. The HSV color space is used for scene category recognition in (Bosch et al. 2008), and the evaluation of local color invariant descriptors is performed in (Burghouts and Geusebroek 2009). The discriminating color space has been discussed in (Liu 2008) and the $I_1I_2I_3$ color space has been shown to possess certain advantages over other color spaces in (Shih and Liu 2005). In this dissertation, eight different color spaces and grayscale have been used for discriminatory feature extraction. These color spaces are discussed in detail in Section 2.1. Fusion of color models, and region and edge detection using color has been investigated for representation of color images (Stokman and Gevers 2007). Some important contributions in color, texture, and shape representation for image retrieval have been discussed in (Datta et al. 2008).

In the last few years, several methods based on Local Binary Patterns (LBP) (Ojala et al. 1994, 1996) features have been proposed for image representation and classification (Zhu et al. 2010; Crosier and Griffin 2008). In a 3×3 neighborhood of an image, the basic LBP operator assigns a binary label 0 or 1 to each surrounding pixel by using the gray value of the central pixel as a threshold and replacing its value with a decimal number converted from the 8-bit binary number. Extraction of LBP features from an image is computationally efficient and with the use of multi-scale filters, partial invariance to rotation and scaling can be achieved (Zhu et al. 2010). Fusion of different types of LBP features, and fusion of LBP and other features have been shown to achieve good image retrieval performance (Crosier and Griffin 2008; Zhang et al. 2007a). Chapter 3 describes scene and texture image classification using new modified LBP operators.

In addition to texture features like LBP, image descriptors based on shape and local features have also been shown to perform well for image retrieval (Zhang et al. 2007a). Several researchers have used the Haar wavelet transform, which represents local features,

for object detection in images (Oren et al. 1997; Papageorgiou et al. 1998; Viola and Jones 2004). Also, LBP has been combined with Haar-like features for face detection (Zhang et al. 2007c). The Histograms of Oriented Gradients (HOG) descriptor (Dalal and Triggs 2005) is a very popular descriptor for representing an image by its local shape, which is captured by the distribution of edge orientations within a region. The Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b) combines the idea of HOG with the Spatial Pyramid Matching (SPM) introduced by (Lazebnik et al. 2006) and stores the spatial distribution of shapes in addition to storing the local shapes. Chapter 4 of this dissertation describes the Haar wavelet transform and the HOG descriptor with greater detail.

In recent times, a lot of researchers have obtained very promising results with part-based methods (Fei-Fei and Perona 2005; Csurka et al. 2004). Here the image is described as a collection of sub-images or regions and the features describe each part and not the whole image. Finally, similar parts are clustered together and a histogram of the parts, rather than the raw features, is used to represent the image. This approach is known as a "bag-of-words model", with each part representing a "visual word" that describes a part of the whole scene (Yang et al. 2007; Jiang et al. 2007). The bag of words model is explained in detail in Chapter 7.

Efficient retrieval requires a robust feature extraction method that is able to extract meaningful low-dimensional patterns from very high dimensional data (Liu 2003). Low-dimensional representation is also important for achieving efficiency in computation. Principal Component analysis (PCA) has been a popular method for performing dimensionality reduction in image indexing and retrieval systems (Liu and Wechsler 2000). Section 2.2 discusses this technique. The Enhanced Fisher Model (EFM) feature extraction method has achieved good success for the task of image representation and retrieval (Liu and Wechsler 2000). This dissertation uses two classification frameworks. One performs EFM feature extraction followed by the Nearest Neighbor (NN) classification method for assigning class labels to test images. This combination, called the EFM-NN classifier henceforth, is ex-

plained in Section 2.3. The other uses a Support Vector Machine classifier (Vapnik 1995) which is discussed in Section 2.4.

2.1 Color Spaces

This section briefly reviews the eight color spaces used to define the proposed descriptors. Perception of color by the human visual system is made possible by specialized retinal cells called cone cells that contain pigments with different spectral sensitivities. The presence of three types of cones in the human eye sensitive to three different spectra results in trichromatic color vision. This is why, any system for representing the full visible color spectrum requires three variables which form a three-dimensional color space. Each color image, therefore, can be split up into three intensity images that are known as color component images or color planes.

The RGB color space, whose three component images represent the red, green, and blue primary colors, is the common tristimulus space for color image representation on a computer. Other color spaces are usually derived from the RGB color space using either linear or nonlinear transformations. The rgb color space, for example, is formed by normalizing the red, green, and blue components in order to reduce the sensitivity of the RGB images to luminance, surface orientation, and other photographic conditions (Gevers et al. 2006):

$$\begin{aligned} r &= R/(R+G+B), \\ g &= G/(R+G+B), \\ b &= B/(R+G+B). \end{aligned} \tag{2.1}$$

Note that for simplicity, R, G, B represent the red, green, and blue pixel values, respectively.

The $\mathrm{I_1I_2I_3}$ color space is defined by the following linear transformation from the

Figure 2.1 An RGB color image, its grayscale image, and the color component images in the RGB, oRGB, rgb, YIQ, HSV, $I_1I_2I_3$, YCbCr and DCS color spaces, respectively.

RGB color space (Ohta 1985):

$$\begin{aligned} I_1 &= (R+G+B)/3, \\ I_2 &= (R-B)/2, \\ I_3 &= (2G-R-B)/4. \end{aligned} \tag{2.2}$$

The HSV (hue, saturation, and value) color space, however, is derived nonlinearly

from the RGB color space (Smith 1978):

$$
\begin{aligned}
H &= \begin{cases} 60(\frac{G-B}{\delta}) & \text{if } MAX = R \\ 60(\frac{B-R}{\delta}+2) & \text{if } MAX = G \\ 60(\frac{R-G}{\delta}+4) & \text{if } MAX = B \end{cases} \\
S &= \begin{cases} \delta/MAX & \text{if } MAX \neq 0 \\ 0 & \text{if } MAX = 0 \end{cases} \\
V &= MAX
\end{aligned}
\tag{2.3}
$$

where $MAX = max(R,G,B)$, $MIN = min(R,G,B)$, and $\delta = MAX - MIN$.

The remaining four color spaces used in this dissertation are, again, transformed from the RGB color space using linear transformations.

The YCbCr color space is defined as follows (Gonzalez and Woods 2008):

$$
\begin{bmatrix} Y \\ Cb \\ Cr \end{bmatrix} = \begin{bmatrix} 16 \\ 128 \\ 128 \end{bmatrix} + \begin{bmatrix} 65.481 & 128.553 & 24.966 \\ -37.797 & -74.203 & 112.000 \\ 112.000 & -93.786 & -18.214 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.4}
$$

The YIQ color space is defined as given below (Shih and Liu 2005):

$$
\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5957 & -0.2745 & -0.3213 \\ 0.2115 & -0.5226 & 0.3111 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.5}
$$

The three component images L, C_1, and C_2 of the oRGB color space are defined as follows (Bratkova et al. 2009):

$$
\begin{bmatrix} L \\ C_1 \\ C_2 \end{bmatrix} = \begin{bmatrix} 0.2990 & 0.5870 & 0.1140 \\ 0.5000 & 0.5000 & -1.0000 \\ 0.8660 & -0.8660 & 0.0000 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.6}
$$

The Discriminating Color Space (DCS) (Liu 2008), is derived from the RGB color space by means of discriminant analysis (Fukunaga 1990). The DCS defines discriminating component images via a linear transformation $W_D \in \mathbb{R}^{3\times3}$ from the RGB color space

$$\begin{bmatrix} D_1 \\ D_2 \\ D_3 \end{bmatrix} = W_D \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{2.7}$$

where D_1, D_2, and D_3 are the values of the discriminating component images in the DCS color space. The transformation matrix $W_D \in \mathbb{R}^{3\times3}$ may be derived through a procedure of discriminant analysis (Fukunaga 1990). Let S_w and S_b be the within-class and the between class scatter matrices of the 3-D pattern vector $\mathscr{X}$ respectively where $S_w, S_b \in \mathbb{R}^{3x3}$. The discriminant analysis procedure derives a projection matrix W_D by maximizing the criterion $J_1 = tr(S_w^{-1}S_b)$ (Fukunaga 1990). This criterion is maximized when W_D^t consists of the eigenvectors of the matrix $S_w^{-1}S_b$ (Fukunaga 1990)

$$S_w^{-1}S_bW_D^t = W_D^t\Delta \tag{2.8}$$

where W_D^t, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1}S_b$, respectively. Figure 2.1 shows a color image, its grayscale image, and its color component images in the RGB, oRGB, rgb, YIQ, HSV, $I_1I_2I_3$, YCbCr and DCS color spaces, respectively. The grayscale image here is an intensity image generated from the RGB image by forming a weighted sum of the R, G, and B components:

$$Gray = 0.2990R + 0.5870G + 0.1140B \tag{2.9}$$

Note that these are the same weights used to compute the Y component of the YIQ color space.

2.2 Principal Component Analysis (PCA)

Principal component analysis, or PCA, which is the optimal feature extraction method in the sense of the mean-square-error, derives the most expressive features for signal and image representation. Specifically, let $\mathscr{X} \in \mathbb{R}^N$ be a random vector whose covariance matrix is defined as follows (Fukunaga 1990):

$$S = \mathscr{E}\{[\mathscr{X} - \mathscr{E}(\mathscr{X})][\mathscr{X} - \mathscr{E}(\mathscr{X})]^t\} \tag{2.10}$$

where $\mathscr{E}(\cdot)$ represents expectation and t the transpose operation. The covariance matrix S is factorized as follows (Fukunaga 1990):

$$S = \Phi \Lambda \Phi^t \tag{2.11}$$

where $\Phi = [\phi_1 \phi_2 \cdots \phi_N]$ is an orthogonal eigenvector matrix and $\Lambda = diag\{\lambda_1, \lambda_2, \ldots, \lambda_N\}$, a diagonal eigenvalue matrix with diagonal elements in decreasing order.

Decorrelation is an important property of PCA, i.e. the components of the transformed data, $\mathscr{X}' = \Phi^t \mathscr{X}$, are decorrelated since the covariance matrix of $\mathscr{X}'$ is diagonal, $\Sigma_{X'} = \Lambda$, and the diagonal elements are the variances of the corresponding components. A second important property of PCA is its optimal signal reconstruction with respect to minimum Mean Square Error (MSE) when just a subset of the principal components is used to represent the original signal. A popular application of this second property is the extraction of the most expressive features of $\mathscr{X}$. Towards that end, a new vector $\mathscr{Y}$ is defined: $\mathscr{Y} = P^t \mathscr{X}$, where $P = [\phi_1 \phi_2 \ldots \phi_K]$, and $K < N$. The most expressive features of $\mathscr{X}$ thus define the new vector $\mathscr{Y} \in \mathbb{R}^K$, which consists of the most significant principal components.

2.3 The Enhanced Fisher Model for Feature Extraction and the Nearest Neighbor Classification Rule — the EFM-NN Classifier

Object and scene image classification using the new descriptors introduced in this dissertation is implemented using the Enhanced Fisher Model (EFM) for feature extraction (Liu and Wechsler 2000) and the Nearest Neighbor (NN) to the mean classification rule for classification. This EFM feature extraction and NN classification procedure is referred to as the EFM-NN classifier throughout this dissertation.

In pattern recognition, a popular method, Fisher's Linear Discriminant (FLD), applies first PCA for dimensionality reduction and then discriminant analysis for feature extraction. PCA is discussed in the previous section, and discriminant analysis often optimizes a criterion defined on the within-class and between-class scatter matrices S_w and S_b, which are defined as follows (Fukunaga 1990):

$$S_w = \sum_{i=1}^{L} P(\omega_i)\mathscr{E}\{(\mathscr{Y} - M_i)(\mathscr{Y} - M_i)^t | \omega_i\} \tag{2.12}$$

$$S_b = \sum_{i=1}^{L} P(\omega_i)(M_i - M)(M_i - M)^t \tag{2.13}$$

where $P(\omega_i)$ is *a priori* probability, ω_i represent the classes, and M_i and M are the means of the classes and the grand mean, respectively. One discriminant analysis criterion is J_1: $J_1 = tr(S_w^{-1} S_b)$, and J_1 is maximized when Ψ contains the eigenvectors of the matrix $S_w^{-1} S_b$ (Fukunaga 1990):

$$S_w^{-1} S_b \Psi = \Psi \Delta \tag{2.14}$$

where Ψ, Δ are the eigenvector and eigenvalue matrices of $S_w^{-1} S_b$, respectively. The discriminating features are defined by projecting the pattern vector $\mathscr{Y}$ onto the eigenvectors of Ψ:

$$\mathscr{Z} = \Psi^t \mathscr{Y} \tag{2.15}$$

$\mathscr{Z}$ thus contains the discriminating features for image classification.

The FLD method, however, often leads to overfitting when implemented in an inappropriate PCA space. To improve the generalization performance of the FLD method, a proper balance between two criteria should be maintained: the energy criterion for adequate image representation and the magnitude criterion for eliminating the small-valued trailing eigenvalues of the within-class scatter matrix (Liu and Wechsler 2000). As a result, the Enhanced Fisher Model (EFM) is developed to improve upon the generalization performance of the FLD method (Liu and Wechsler 2000). Specifically, the EFM method improves the generalization capability of the FLD method by decomposing the FLD procedure into a simultaneous diagonalization of the within-class and between-class scatter matrices (Liu and Wechsler 2000). The simultaneous diagonalization reveals that during whitening the eigenvalues of the within-class scatter matrix appear in the denominator. Since the small eigenvalues tend to encode noise (Liu and Wechsler 2000), they cause the whitening step to fit for misleading variations, and this leads to poor generalization performance. To enhance performance, the EFM method preserves a proper balance between the need that the selected eigenvalues account for most of the spectral energy of the raw data (for representational adequacy), and the requirement that the eigenvalues of the within-class scatter matrix (in the reduced PCA space) are not too small (for better generalization performance) (Liu and Wechsler 2000).

2.4 Support Vector Machine

The Support Vector Machine (SVM) is a particular realization of statistical learning theory. The approach described by SVM, known as structural risk minimization, minimizes the risk functional in terms of both the empirical risk and the confidence interval (Vapnik 1995). SVM is built from two ideas: (i) a nonlinear mapping of the input space to a high-dimensional feature space, and (ii) designing the optimal hyperplane in terms of the maximal margin between the patterns of the two classes in the feature space. SVM is very

popular and has been applied extensively for pattern classification, regression, and density estimation since it displays a good generalization performance, .

Let $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_k, y_k), \mathbf{x}_i \in \mathbb{R}^N$, and $y_i \in \{+1, -1\}$ be k training samples in the input space, where y_i indicates the class membership of $\mathbf{x}_i$. Let ϕ be a nonlinear mapping between the input space and the feature space, $\phi : \mathbb{R}^N \rightarrow \mathscr{F}$, i.e., $\mathbf{x} \rightarrow \phi(\mathbf{x})$. The optimal hyperplane in the feature space is defined as follows:

$$w_0 \cdot \phi(\mathbf{x}) + b_0 = 0 \tag{2.16}$$

It can be proven (Vapnik 1995) that the weight vector w_0 is a linear combination of the support vectors, which are the vectors $\mathbf{x}_i$ that satisfy $y_i(w_0 \cdot \phi(\mathbf{x}_i) + b_0) = 1$:

$$w_0 = \sum_{supportvectors} y_i \alpha_i \phi(\mathbf{x}_i) \tag{2.17}$$

where α_i's are determined by maximizing the following functional:

$$L(\alpha) = \sum_{i=1}^{k} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{k} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.18}$$

subject to the following constraints:

$$\sum_{i=1}^{k} \alpha_i y_i = 0, \alpha_i \geq 0, i = 1, 2, \ldots, k \tag{2.19}$$

From Equations 2.16 and 2.17, the linear decision function in the feature space can be derived

$$f(\mathbf{x}) = sign(\sum_{supportvectors} y_i \alpha_i \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}) + b_0) \tag{2.20}$$

It should be noted that the decision function (see Equation 2.20) is defined by the dot products in the high dimensional feature space, where computation might be prohibitively expensive. SVM, however, manages to compute the dot products by means of a kernel

function (Vapnik 1995)

$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j) \tag{2.21}$$

Three classes of kernel functions widely used in kernel classifiers, neural networks, and SVMs are polynomial kernels, Gaussian kernels, and sigmoid kernels (Vapnik 1995):

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j)^d, \tag{2.22}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = exp(\frac{-(\|\mathbf{x}_i - \mathbf{x}_j\|)^2}{2\sigma^2}), \tag{2.23}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = tanh(k(\mathbf{x}_i \cdot \mathbf{x}_j) + v), \tag{2.24}$$

where $d \in \mathbb{N}$, $\sigma > 0$, $k > 0$, and $v < 0$.

The SVM implementation used for the experiments presented in this dissertation is the one that is distributed with the VlFeat package (Vedaldi and Fulkerson 2010). The parameters of the support vector machine are tuned empirically using only the training data, and the parameters that yield the best average precision on the training data are used for classification of the test data. In particular, the cost parameter C has been empirically set to 100 for the best classification performance in the experiments described here.

CHAPTER 3

NOVEL MULTI-MASK LBP (MLBP) IMAGE DESCRIPTORS

This chapter first discusses the traditional Local Binary Patterns (LBP) feature for texture representation. Then, the multi-mask LBP (mLBP) is introduced. Next, the mLBP descriptors from different color spaces are fused to form the novel Color LBP Fusion (CLF) and the Color Grayscale LBP Fusion (CGLF) descriptors. Finally, the effectiveness of the proposed method is demonstrated using two texture image datasets and one scene image dataset.

3.1 Local Binary Patterns (LBP)

The Local Binary Patterns (LBP) method derives the texture description of a grayscale i.e. intensity image by comparing a center pixel with its neighbors (Ojala et al. 1994, 1996, 2002). In particular, for a 3×3 neighborhood of a pixel $\mathbf{p} = [x, y]^t$, $\mathbf{p}$ is the center pixel used as a threshold. The neighbors of the pixel $\mathbf{p}$ are defined as $N(\mathbf{p}, i) = [x_i, y_i]^t$, $i = 0, 1, \cdots, 7$, where i is the number used to label the neighbor. The value of the LBP code of the center pixel $\mathbf{p}$ is calculated as follows:

$$LBP(\mathbf{p}) = \sum_{i=0}^{7} 2^i S\{G[N(\mathbf{p}, i)] - G(\mathbf{p})\} \tag{3.1}$$

where $G(\mathbf{p})$ and $G[N(\mathbf{p}, i)]$ are the gray levels of the pixel $\mathbf{p}$ and its neighbor $N(\mathbf{p}, i)$, respectively. S is a threshold function that is defined below:

$$S(x_i - x_c) = \begin{cases} 1, & \text{if } x_i \geq x_c \\ 0, & \text{otherwise} \end{cases} \tag{3.2}$$

LBP tends to achieve grayscale invariance because only the signs of the differences

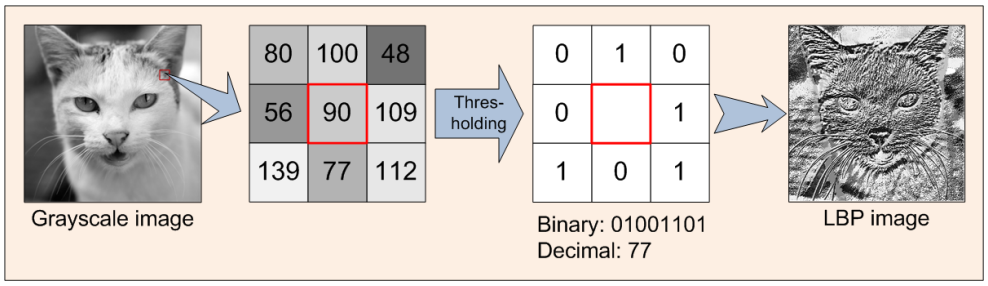


Figure 3.1 A grayscale image, its LBP image, and the illustration of the computation of the LBP code for a center pixel with gray level 90.

between the center pixel and its neighbors are used to define the value of the LBP code as shown in Equation 3.1. Figure 3.1 shows a grayscale image on the left and its LBP image on the right. The two 3×3 matrices in the middle illustrate how the LBP code is computed for the center pixel whose gray level is 90. In particular, the center pixel functions as a threshold, and after thresholding the right 3×3 matrix reveals the signs of the differences between the center pixel and its neighbors. Note that the signs are derived from Equations 3.1 and 3.2, and the threshold value is 90, as the center pixel is used as the threshold in the LBP definition. The binary LBP code is 01001101, which corresponds to 77 in decimal.

3.2 A New Color Multi-mask LBP (mLBP)

The traditional LBP descriptor described in the preceding section assigns an intensity value to each pixel of an image based on the intensity values of just eight pixels adjoining it. The new mLBP feature discussed here is generated by comparing the value of each pixel to its eight neighbors in the three neighborhoods shown in the top row of Figure 3.2. In particular, the first neighborhood generates the traditional LBP image shown in the first column of the second row in Figure 3.2. The second neighborhood is a square rotated by 45° which produces the image shown in the second column of the second row, and the third neighborhood chooses eight pixels farther away from the center pixel, thus producing the image

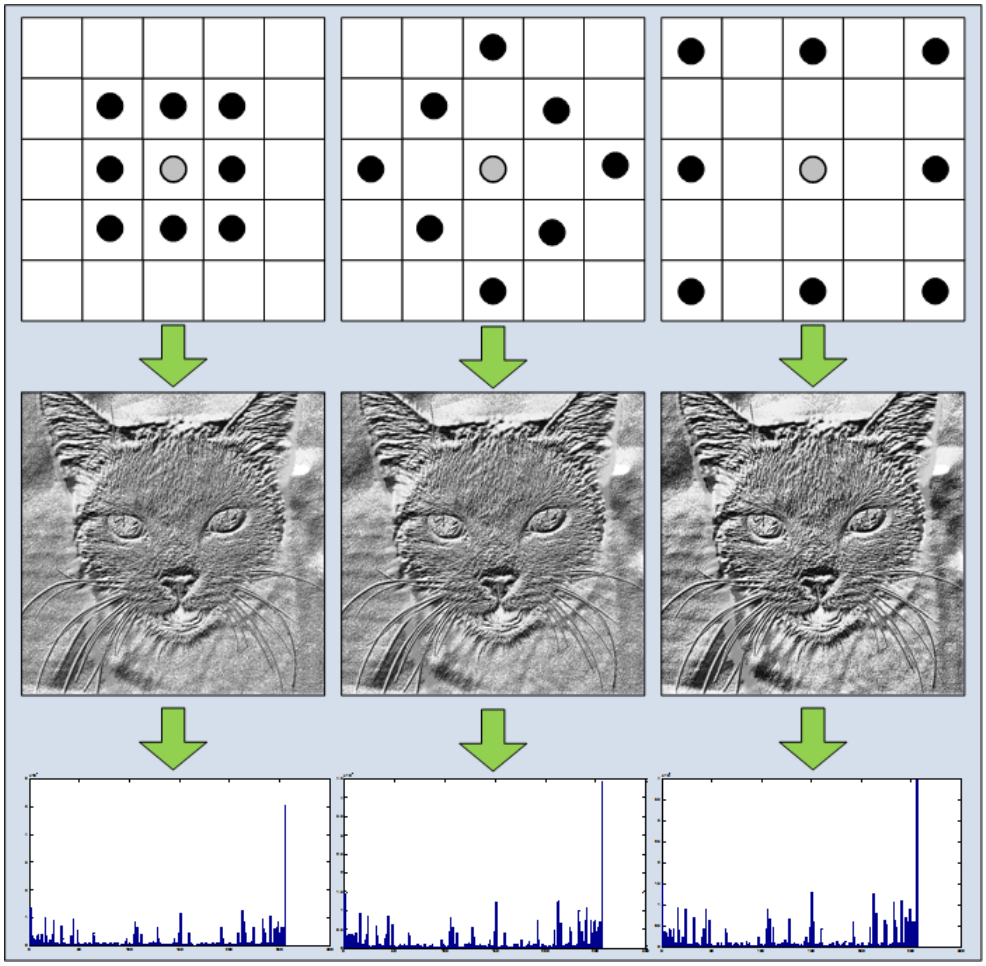


Figure 3.2 The three neighborhood for extracting the multi-mask LBP (mLBP) descriptor, the three LBP images generated from the three neighborhoods, and the three histograms generated from these images. The original image is the same as used in Figure 3.1.

shown in the third column of the second row. By choosing these three different neighborhoods, partial invariance to scaling and rotation can be achieved. The histograms from the three LBP images shown in the bottom row of Figure 3.2 are concatenated and used as a feature vector which is independent of the image size. The three 256-bin histograms when concatenated generate a 768 feature vector for a grayscale image.

To encode the variations in intensity and texture present in the different color component images, the color mLBP descriptor is derived by individually computing the mLBP descriptor on each of the three component images in the specific color space. This produces

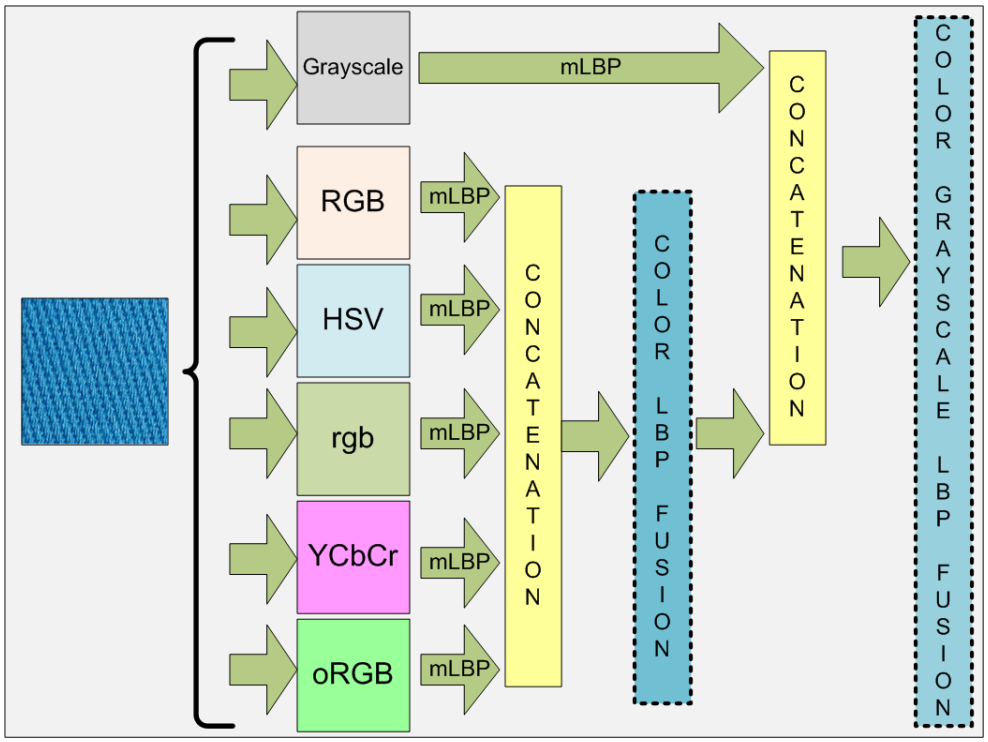


Figure 3.3 A schematic diagram showing the generation of the CLF and the CGLF descriptors from a color image taken from the cotton category of the KTH-TIPS2-b dataset.

a 2304 dimensional descriptor that is formed from concatenating the 768 dimensional vectors from the three channels. In particular, the RGB-mLBP, the HSV-mLBP, the YCbCr-mLBP, the rgb-mLBP and the oRGB-mLBP descriptors are constructed by concatenating the mLBP descriptors of the three component images in the RGB, HSV, YCbCr, rgb and oRGB color spaces respectively.

3.3 Fusing the Multi-mask LBP Descriptor in Different Color Spaces and Grayscale

As will be seen from the experiments in Section 3.4, the color multi-mask LBP descriptor yields different classification rates for different color spaces. This indicates that the information contained by the different color mLBP descriptors is not the same, and if combined, these descriptors could classify more accurately. Hence, the Color LBP Fusion

(CLF) descriptor is formed by fusing the RGB-mLBP, the YCbCr-mLBP, the HSV-mLBP, the oRGB-mLBP, and the rgb-mLBP descriptors. Further, the Color Grayscale LBP Fusion (CGLF) descriptor is obtained by fusing the CLF descriptor and the grayscale-mLBP descriptor. Indeed, it will be seen in Section 3.4.2 that these two descriptors achieve a better classification performance than the mLBP descriptors in the individual color spaces. Figure 3.3 shows the process by which the CLF and the CGLF descriptors are generated. Specifically, it shows a color image (on the left) being converted to five different color spaces and grayscale. Each of these converted images then undergo an mLBP process. The descriptors resulting from the five color spaces are concatenated to form the CLF feature vector. The CLF is further fused with the grayscale mLBP features to get the CGLF feature vector. These two feature vectors are represented using blue rectangles with broken-line borders in the figure. The results of the classification experiments with the different new descriptors are described and evaluated in the next section.

3.4 Experiments

Three publicly available and fairly challenging image datasets are used to evaluate the proposed descriptors applying the EFM-NN classification method. This section first briefly describes the three datasets used. Then a comparative assessment of the classification performance of the mLBP descriptor is done in five different color spaces and grayscale. Next the performance of the two fused descriptors - CLF and CGLF - is evaluated, and the classification performance is compared with the results of other researchers.

3.4.1 Datasets Used

The three image datasets used for evaluating the mLBP desciptor are briefly introduced in this section. All of these datasets are widely used for evaluating the performance of texture and scene image descriptors and classification methods.

The MIT Scene Dataset

The MIT Scene dataset (also known as OT Scenes) (Oliva and Torralba 2001) has 2,688 images classified as eight categories: 360 coast, 328 forest, 260 highway, 308 inside of cities, 374 mountain, 410 open country, 292 streets, and 356 tall buildings. All of the images are in color, in JPEG format, and the size of each image is 256×256 pixels. There is a large variation in light, content and angles, along with a high intra-class variation. The sources of the images vary (from commercial databases, websites, and digital cameras) (Oliva and Torralba 2001). Figure 3.4(a) shows some images from this dataset.

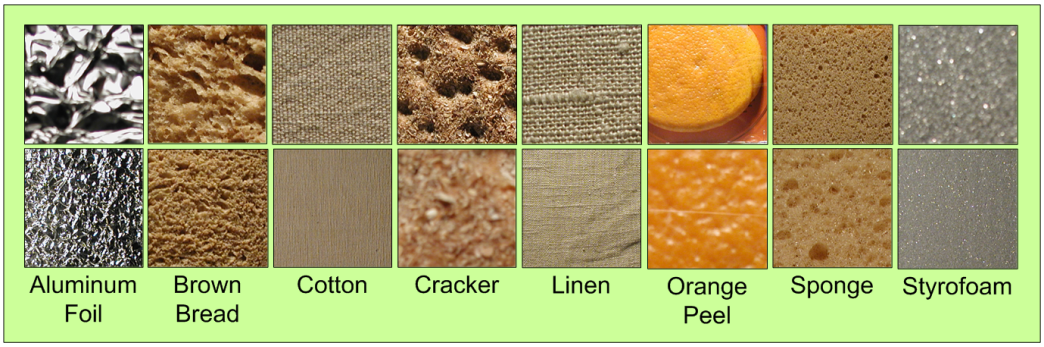
The KTH-TIPS Dataset

The KTH-TIPS (Textures under varying Illumination, Pose and Scale) dataset (Hayman et al. 2004; Caputo et al. 2005; Kondra and Torre 2008) consists of 10 classes of textures with 81 images per class. All the images are in color, PNG format and the maximum image size is 200x200 pixels. All ten textures have been photographed at nine scales and nine illumination conditions for each scale. Some of the classes have a very similar visual appearance, like cotton and linen, and brown bread and sponge which makes this dataset moderately challenging. See Figure 3.4(b) for some sample images from this dataset.

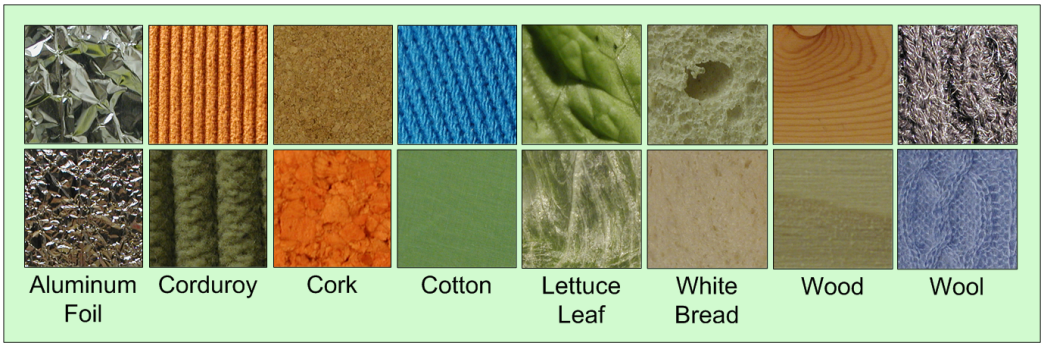
The KTH-TIPS2-b Dataset

The KTH-TIPS2-b dataset (Caputo et al. 2005) is a more challenging extension of the KTH-TIPS dataset with 11 classes of materials and 4 samples for each material. Each of these samples has 108 images with 432 images per class and a total of 4752 images. Some of the images in the classes like wool and cotton are from differently colored samples leading to very high intra-class variation among samples, while some samples from different classes like cork and cracker have the same color and general appearance thus lowering the inter-class variation. See Figure 3.4(c) for some sample images from this dataset.

(a)



(b)



(c)

Figure 3.4 Some sample images from (a) the MIT Scene dataset, (b) the KTH-TIPS dataset, and (c) the KTH-TIPS2-b dataset. Please note that all texture categories from the KTH-TIPS and KTH-TIPS2-b datasets are not shown in the figure.

3.4.2 Comparative Assessment of the Multi-mask LBP Descriptor in Different Color Spaces and Grayscale

The mLBP descriptor is now assessed in grayscale and five different color spaces — the RGB, oRGB, HSV, YCbCr, and rgb color spaces — for image classification performance using the three datasets. To derive the mLBP descriptor from each color image, the mLBP

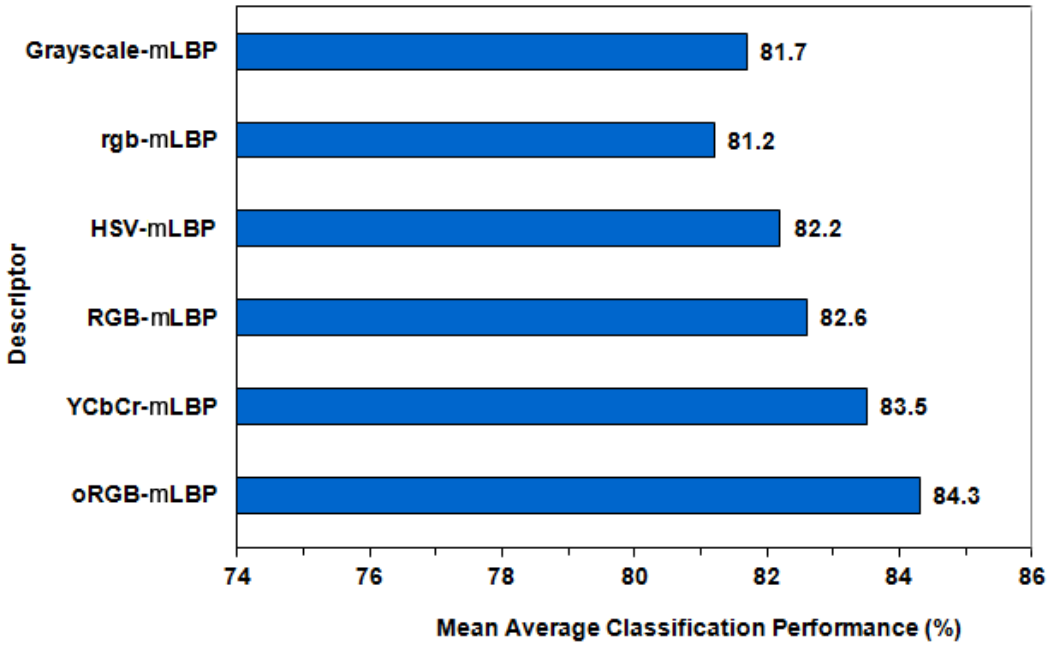


Figure 3.5 The mean average classification performance of the six descriptors using the EFM classifier on the MIT scene dataset: the grayscale-mLBP, the rgb-mLBP, the HSV-mLBP, the RGB-mLBP, the YCbCr-mLBP and the oRGB-mLBP.

descriptor is computed from each color component and concatenated. Each image is transformed in the five color spaces and the same operations are performed to construct the five different color mLBP descriptors. Each color image is also converted to grayscale and its mLBP descriptor is extracted. Next PCA is applied to reduce the dimensionality of the mLBP descriptors to derive the most expressive features, which are further processed by EFM to obtain the most discriminatory features for classification, and the nearest neighbor rule is finally used for image classification. The classification task is to assign each test image to one of a number of categories. The performance is measured using a confusion matrix, and the overall performance rates are measured by the average value of the diagonal entries of the confusion matrix.

For the MIT scene dataset, five image sets are randomly selected. Each set consists of 2000 images for training (250 images per class) and the rest 688 images for testing. The

first set of experiments assesses the overall classification performance of the six descriptors. Note that for each category five-fold cross validation is implemented for each descriptor to derive the average classification performance. As a result, each descriptor yields eight average classification rates corresponding to the eight image categories. The mean value of these eight average classification rates is defined as the mean average classification performance for the descriptor. Figure 3.5 shows the mean average classification performance of various mLBP descriptors. Specifically, the vertical axis shows the different descriptors and the horizontal axis shows the mean average classification performance. The best recognition rate that is obtained is for the oRGB-mLBP which achieves the classification rate of 84.3%. The YCbCr-mLBP, RGB-mLBP, HSV-mLBP, rgb-mLBP and grayscale-mLBP yield classification rates of 83.5%, 82.6%, 82.2% 81.2% and 81.7%, respectively.

The second set of experiments assesses the six descriptors using the EFM-NN classifier on individual image categories. From Table 3.1 it can be seen that for oRGB-mLBP, the top seven categories achieve a success rate of 80% or more. The Forest category achieves a success rate of over 94% across all six descriptors. Individual color mLBP descriptors improve upon the grayscale-mLBP on most of the categories. The classification performance of the fused descriptors CLF and CGLF which are discussed in Section 3.4.3 have also been included in this table for comparison.

For the KTH-TIPS dataset, five random sets of 40 training images per class and 41 test images per class are selected (same numbers as used in (Crosier and Griffin 2008; Zhang et al. 2007a; Kondra and Torre 2008)). Within each set there is no overlap in the images selected for training and testing. Since this dataset is moderately challenging, only the oRGB-mLBP is tested on this dataset which gives a classification performance of 99.1%. The rest of the detailed experiments are done on the KTH-TIPS2-b dataset which is a more complex and extended version of this dataset.

For the KTH-TIPS2-b dataset, five random sets of 200 training images per class and 100 testing images per class are used with no common images between the training

Table 3.1 Category Wise Descriptor Performance (%) Split-out with the EFM-NN Classifier on the MIT Scene Dataset. Note that the Categories are Sorted on the CGLF Results

Category	CGLF	CLF	oRGB mLBP	YCbCr mLBP	RGB mLBP	HSV mLBP	rgb mLBP	Gray mLBP
Forest	**97**	**97**	**97**	**97**	95	94	94	94
Highway	90	**93**	90	87	90	90	90	**93**
Street	**90**	86	83	83	82	84	82	81
Coast	**88**	87	85	**88**	83	81	82	86
Inside City	**87**	**87**	86	83	81	80	79	83
Tall Building	**86**	**86**	**86**	83	84	82	80	79
Mountain	**85**	84	80	81	80	80	76	77
Open Country	**71**	**71**	68	66	65	66	68	61
Mean	**86.6**	**86.4**	**84.2**	**83.5**	**82.6**	**82.2**	**81.2**	**81.7**

and test sets. The first set of experiments assesses the overall classification performance of the six descriptors on the KTH-TIPS2-b dataset. Note that for each category five-fold cross validation is implemented for each descriptor using the EFM-NN classifier to derive the average classification performance. Figure 3.6 shows the mean average classification

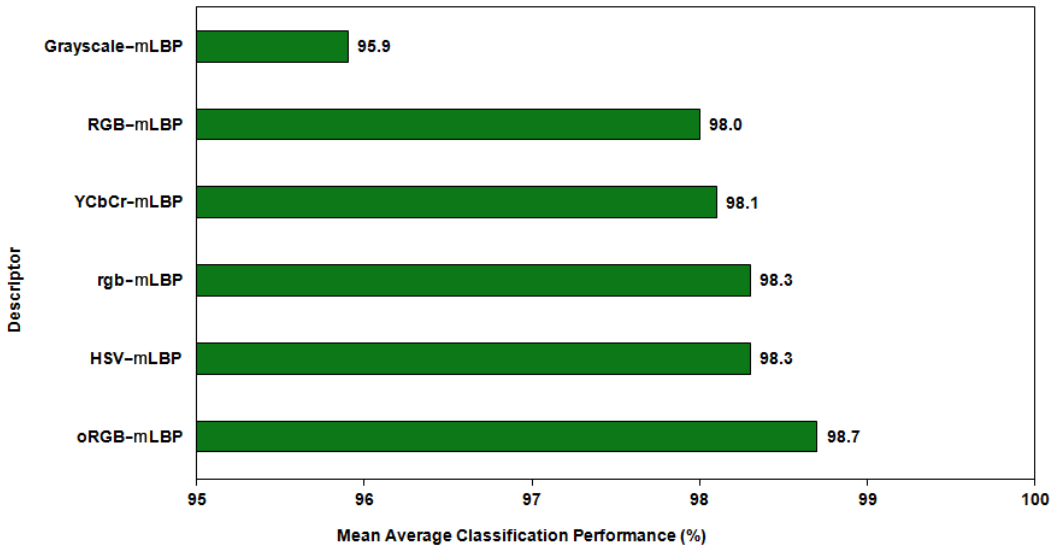


Figure 3.6 The mean average classification performance of the six descriptors using the EFM classifier on the KTH-TIPS2-b dataset: the grayscale-mLBP, the RGB-mLBP, the YCbCr-mLBP, the rgb-mLBP, the HSV-mLBP, and the oRGB-mLBP descriptors.

Table 3.2 Category Wise Descriptor Performance (%) Split-out with the EFM-NN Classifier on the KTH-TIPS2-b Dataset. Note that the Categories are Sorted on the CGLF Results

Category	**CGLF**	**CLF**	**oRGB mLBP**	**HSV mLBP**	**rgb mLBP**	**Gray mLBP**
Aluminum Foil	**100**	**100**	**100**	**100**	**100**	**100**
Brown Bread	**100**	**100**	**100**	99	99	94
Corduroy	**100**	**100**	**100**	**100**	**100**	93
Cork	**100**	**100**	**100**	98	98	98
Cracker	**100**	**100**	96	93	93	90
Lettuce Leaf	**100**	**100**	**100**	**100**	**100**	97
Linen	**100**	**100**	**100**	99	99	99
Wood	**100**	**100**	**100**	**100**	**100**	**100**
Wool	**100**	**100**	99	**100**	**100**	96
White Bread	**99**	**99**	**99**	**99**	**99**	97
Cotton	**98**	97	97	96	96	91
Mean	**99.6**	**99.6**	**98.7**	**98.3**	**98.3**	**95.9**

performance of various descriptors. The best recognition rate that is obtained is 98.7% for the oRGB-mLBP, followed by the HSV-mLBP, rgb-mLBP, YCbCr-mLBP, RGB-mLBP and grayscale-mLBP with 98.3%, 98.3%, 98.1%, 98.0% and 95.9% success rates, respectively.

The second set of experiments assesses the three best color mLBP descriptors and the grayscale-mLBP using the EFM classifier on individual image categories. From Table 3.2 it can be seen that for the oRGB-mLBP descriptor, seven out of eleven categories achieve 100% success rate and all of the categories achieve a success rate of 96% or more. Aluminum Foil, Corduroy, Lettuce Leaf and Wood achieve 100% success rate across the best three descriptors. Individual color mLBP descriptors improve upon the grayscale-mLBP on most of the categories. Here also, the classification performance of the fused descriptors CLF and CGLF which are discussed in Section 3.4.3 have been included in this table for comparison.

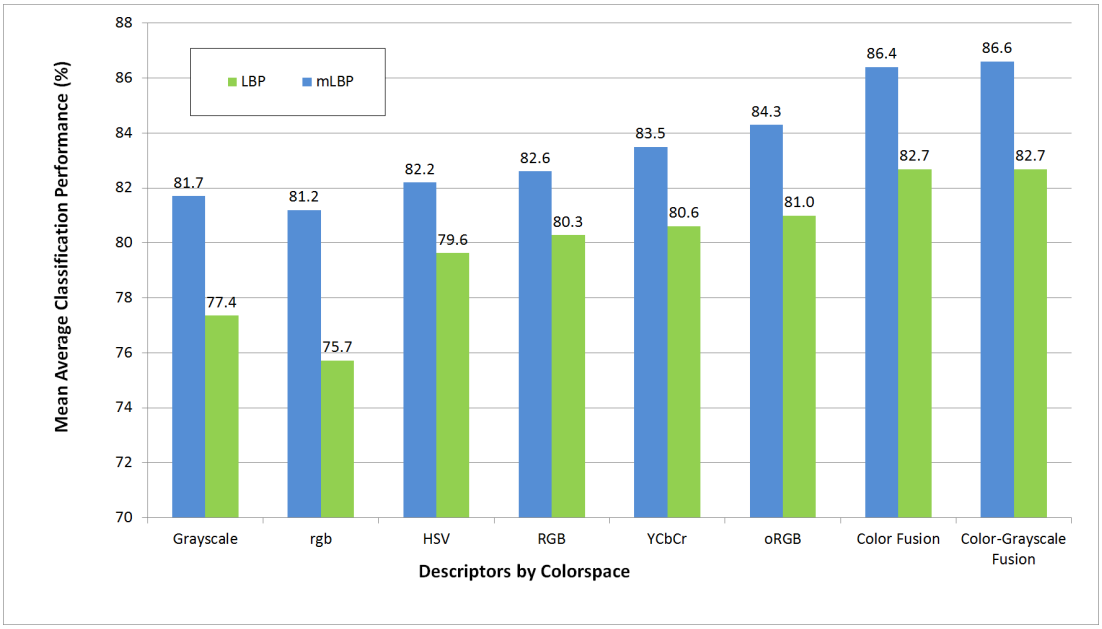


Figure 3.7 The mean average classification performance of the six single-color mLBP descriptors and the two fused-color mLBP descriptors on the MIT scene dataset compared with the classification performance of the traditional LBP descriptors in the same color spaces.

3.4.3 Comparative Assessment of the Proposed Descriptors and Some Popular Image Descriptors

In this section, the performance of the proposed CLF and CGLF descriptors are evaluated on the three datasets described in Section 3.4.1. Then the CLF and CGLF descriptors are compared with some other popular descriptors for classification performance. The mLBP descriptors in the individual color spaces and grayscale, as well as the fused mLBP descriptors (CLF and CGLF) are also compared with the traditional LBP descriptors in the respective color spaces.

On the MIT Scene dataset, the CLF achieves a classification performance of 86.4% which is 2.1% more than the best single-color descriptor performance. The CGLF outperforms this slightly, yielding a 86.6% success rate. As can be seen from Table 3.1, the CLF results on each of the eight categories show significant improvement upon the grayscale-mLBP and the CGLF slightly improves upon the CLF. It should be noted that fusion of the

Table 3.3 Comparison of the Classification Performance (%) with other Methods on the MIT Scene Dataset

#train	#test	Proposed Method		LBP	PHOG
		CLF	**86.4**		
2000	688	CGLF	**86.6**	82.7	79.1
		CLF	**80.3**		
800	1888	CGLF	**80.3**	77.9	-

color mLBP descriptors (CLF) improves upon the grayscale-mLBP by a significant 4.7% margin.

Figure 3.7 shows the comparative classification performance of the proposed mLBP descriptor and the traditional single-mask LBP descriptor. Specifically, it compares the classification performance achieved by the mLBP descriptor and the LBP descriptor in the rgb, HSV, RGB, YCbCr, oRGB color spaces and grayscale. It also fuses the LBP descriptors from different color spaces and grayscale and compares their performance with CLF and CGLF. The horizontal axis shows the different descriptors and the vertical axis shows the classification performance as a percentage. As can be seen from the figure, using three neighborhoods significantly improves classification performance in all color spaces as well as in grayscale and fusion of color spaces. Table 3.3 compares classification results obtained using the proposed descriptors with that obtained by using traditional LBP as described by (Ojala et al. 1994), and also with that obtained by applying the Pyramid of Histograms of Oriented Gradients (PHOG) as described by (Bosch et al. 2007b) to this dataset.

It may be noted here that fusing PHOG with CGLF produces a classification performance of 89.5% on the MIT Scene dataset with 250 training images per class and 84.3% with 100 training images per class, results which are higher than (Oliva and Torralba 2001), and this was further explored in (Banerji et al. 2011). However, that detail is not included here as PHOG is not an original work proposed in this dissertation and hence this dissertation does not focus on PHOG and its fusion methods.

Table 3.4 Comparison of the Classification Performance (%) with Other Methods on the KTH-TIPS Dataset

Methods	Performance
Proposed Method:	
CGLF	**99.6**
CLF	**99.6**
(Crosier and Griffin 2008)	98.5
(Kondra and Torre 2008)	97.7
(Zhang et al. 2007a)	95.5

Classification experiments using the CLF and CGLF descriptors were run on the KTH-TIPS dataset with the aforementioned training and test image sets. For this dataset, the CLF and the CGLF descriptors are tied at 99.6%. Table 3.4 shows a comparison of results from the proposed descriptors with those obtained from other methods in (Crosier and Griffin 2008; Zhang et al. 2007a; Kondra and Torre 2008). Combined mLBP descriptors (CLF and CGLF) improve upon the result in (Crosier and Griffin 2008), previously the best result on this dataset.

For the KTH-TIPS2-b dataset also, the CLF generates a success rate of 99.6% and CGLF very slightly improves upon the CLF. This, however, does not necessarily indicate that the grayscale information is redundant since it can be seen from Table 3.2 that almost all the categories show a success rate of 100% with these two descriptors. It only indicates that CLF alone contains enough information to correctly classify the texture images in the case of KTH-TIPS2-b dataset.

3.5 Summary

Three new color descriptors have been proposed in this chapter: the oRGB-mLBP descriptor, the Color LBP Fusion (CLF), and the Color Grayscale LBP Fusion (CGLF) descriptors for scene image and texture image classification with applications to image search and retrieval. Results of the experiments using three popular datasets show that the oRGB-mLBP

descriptor's recognition performance is better than other color mLBP descriptors, and the CLF and the CGLF descriptors perform better than the single color mLBP descriptors. The fusion of multiple Color LBP descriptors (CLF) and Color Grayscale LBP descriptor (CGLF) show improvement in the classification performance, which indicates that various color mLBP descriptors are not redundant for image classification tasks.

CHAPTER 4

HAARHOG: IMPROVING THE HOG DESCRIPTOR BY ENHANCING LOCAL FEATURES

Chapter 3 discusses the LBP descriptor and introduces a novel mLBP descriptor to encode the texture of an image. Apart from texture, shape and high-frequency local information also contribute heavily to object and scene image recognition, and hence, descriptors based on such features are frequently used for image classification. The Histograms of Oriented Gradients (HOG) descriptor (Dalal and Triggs 2005), which represents an image by histograms of the pixel gradients at the object edges in an image, stores information about the shapes contained in the image. As a result, HOG has become a popular descriptor for object detection and content based image retrieval. Wavelets are known to selectively enhance high frequency local information in selected orientations. That is why wavelets, such as the Haar wavelets have been widely applied for object detection in images (Zhang et al. 2007c).

This chapter introduces a novel image descriptor based on shape and local high-frequency features from an image, and then extends it to include the benefits of using multiple color spaces. Specifically, first, a new HaarHOG feature vector is defined that extracts shape as well as other local features from a grayscale image by combining the Haar wavelet transform with the Histograms of Oriented Gradients (HOG). Next, the definition of the new descriptor is extended for use on color images.

To assess the classification performance of the proposed descriptor, a Support Vector Machine (SVM) classifier with a linear kernel is used on several widely used and publicly available image datasets. In these experiments, it is shown to achieve a significantly better classification performance than the conventional HOG descriptor, as well as some other popular image descriptors, such as Scale Invariant Feature Transform (SIFT) based methods, Spatial Envelope (SE), Object Bank (OB), as well as Local Binary Patterns (LBP).

4.1 Haar Wavelet Transform

The 2D Haar wavelet transform is defined as the projection of an image onto the 2D Haar basis functions, which are formed by the tensor product of the one dimensional Haar scaling and wavelet functions (Burrus et al. 1998; Beylkin et al. 1991). The Haar scaling function $\phi(x)$ is defined below (Burrus et al. 1998; Porwik and Lisowska 2004):

$$\phi(x) = \begin{cases} 1, & 0 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

A family of functions can be generated from the basic scaling function by scaling and translation (Burrus et al. 1998; Porwik and Lisowska 2004):

$$\phi_{i,j}(x) = 2^{i/2}\phi(2^i x - j) \tag{4.2}$$

As a result, the scaling functions $\phi_{i,j}(x)$ can span the vector spaces V^i, which are nested as follows: $V^0 \subset V^1 \subset V^2 \subset \cdots$ (Mallat 1989).

The Haar wavelet function $\psi(x)$ is defined as follows (Burrus et al. 1998; Porwik

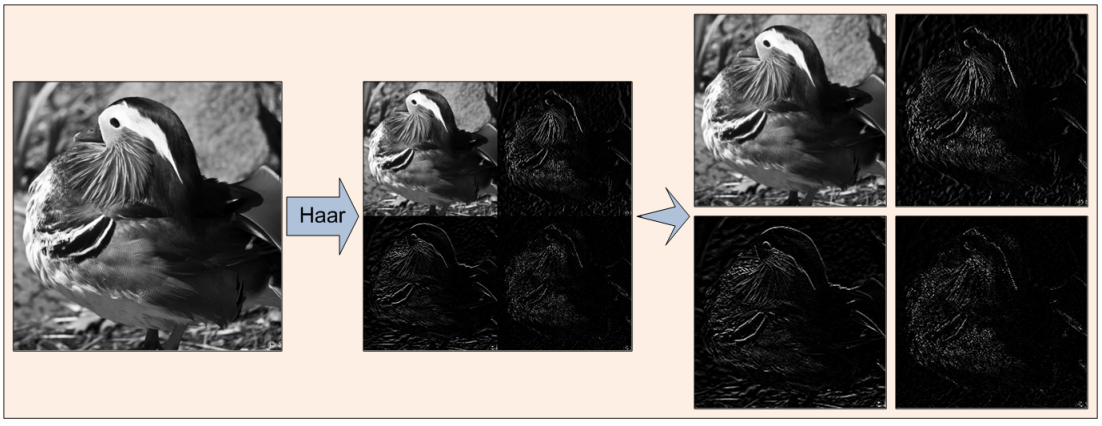


Figure 4.1 A grayscale image and its Haar wavelet transform.

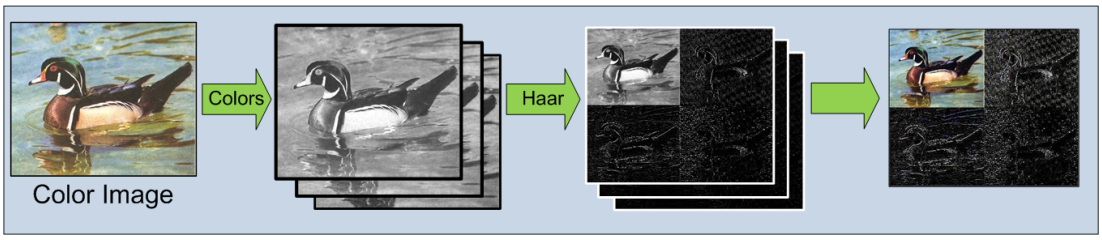


Figure 4.2 A color image taken from the duck category of the Caltech 256 dataset, its three color component images, their Haar wavelet transformed images, and the color Haar wavelet transformed image.

and Lisowska 2004):

$$\psi(x) = \begin{cases} 1, & 0 \leq x < 1/2 \\ -1, & 1/2 \leq x < 1 \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

The Haar wavelets are generated from the mother wavelet by scaling and translation (Burrus et al. 1998; Porwik and Lisowska 2004):

$$\psi_{i,j}(x) = 2^{i/2}\psi(2^i x - j) \tag{4.4}$$

The Haar wavelets $\psi_{i,j}(x)$ span the vector space W^i, which is the orthogonal complement of V^i in V^{i+1}: $V^{i+1} = V^i \oplus W^i$ (Burrus et al. 1998; Porwik and Lisowska 2004). The 2D Haar basis functions are the tensor product of the one dimensional scaling and wavelet functions (Beylkin et al. 1991).

Figure 4.1 shows a grayscale image of a Mandarin duck and its Haar wavelet transformed image. The right side of the figure displays an enlargement of the four quadrants of the Haar wavelet transformed image which shows that different sub-images enhance high-frequency local features in different orientations. Figure 4.2 shows the generation of the Haar wavelet transformed image for a color image with three component intensity images. Specifically, a color image is split into its three component planes and then the Haar wavelet transformation is applied to each plane separately to get the color Haar transformed image

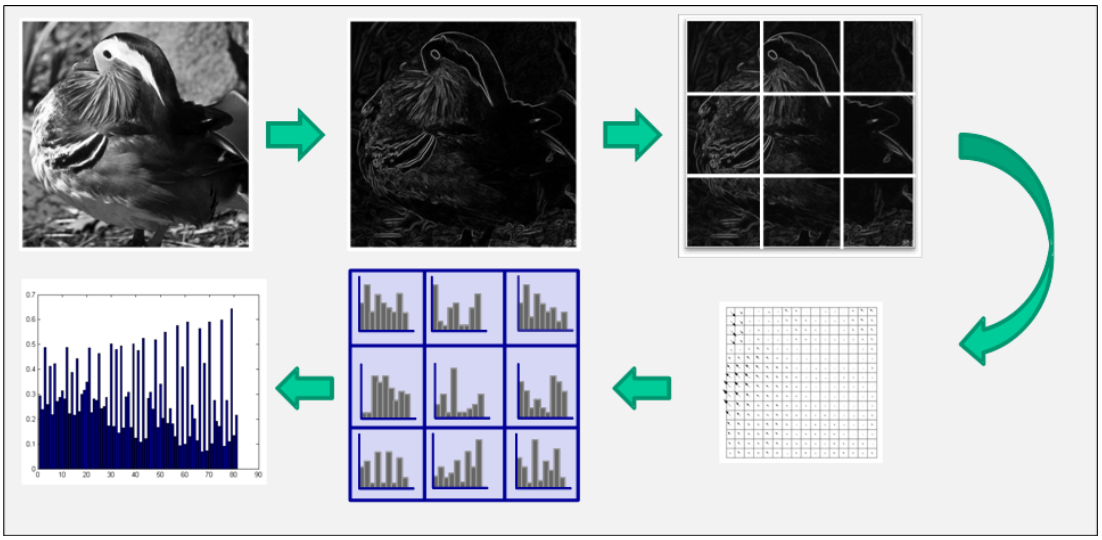


Figure 4.3 A grayscale image and the formation of the HOG descriptor.

with four color sub-images.

4.2 Histogram of Oriented Gradients (HOG)

The idea of Histogram of Oriented Gradients (HOG) rests on the observation that local features such as object appearance and shape can often be characterized well by the distribution of local intensity gradients in the image (Dalal and Triggs 2005). HOG features are derived from an image based on a series of normalized local histograms of image gradient orientations in a dense grid (Dalal and Triggs 2005; Ludwig et al. 2009). In particular, first the gradient magnitude and direction is calculated at each pixel in the image. The image window is then divided into blocks and the blocks into small cells. For each cell, a local histogram of the gradient directions weighted by the gradient magnitudes is accumulated over all the pixels of the cell. All the histograms within a block of cells are then normalized to reduce the effect of illumination variations. The blocks can be overlapped with each other for performance improvement. The final HOG descriptors are formed by concatenating the normalized histograms from all the blocks into a single vector.

Figure 4.3 demonstrates the formation of the HOG vector for a grayscale image.

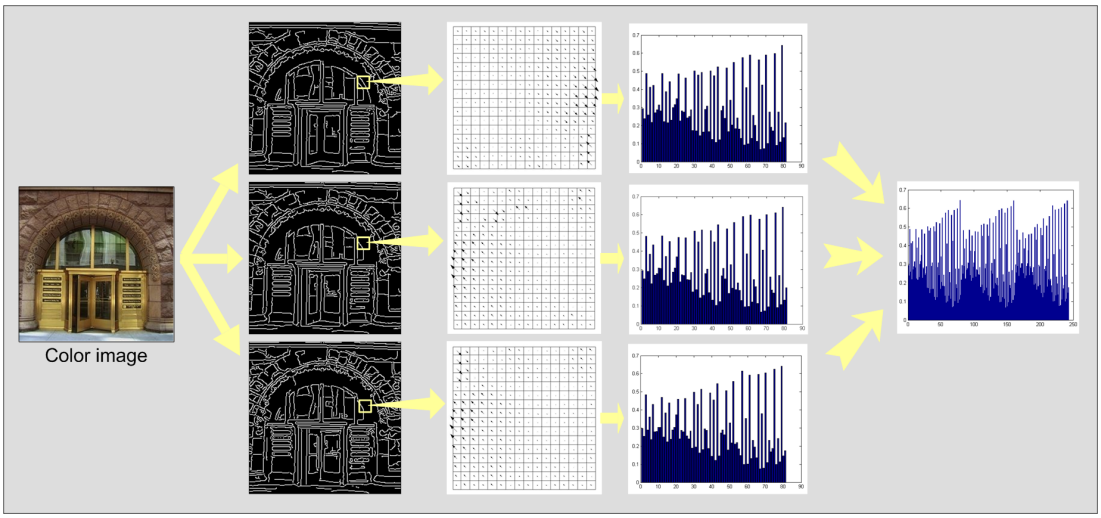


Figure 4.4 A color image taken from the inside-city category of the MIT Scene dataset, the gradient magnitude images of its three color components, the orientation gradients of an example small area from every gradient magnitude image, the three HOG descriptors for the three color component images, and the concatenated HOG descriptor for the whole color image.

The image of a duck at the top right is the original grayscale image. The first step is the calculation of the gradient magnitudes at every pixel. The gradient magnitude image is shown in the middle figure of the top row. This resembles an edge image as images usually have high gradient magnitudes near the edges. Next, the image window is divided into a number of blocks as shown in the last image in the first row of Figure 4.3. In the original implementation by (Dalal and Triggs 2005), dividing the image into 3×3 blocks as shown in the figure was found to be optimal for pedestrian detection. For the experiments presented here, however, the performance was found to be increasing up to 5×5 blocks and so 5×5 blocks were used for this implementation. Next, the orientation of each pixel in each block is put in one of 10 orientation bins weighted by its magnitude and thus a weighted histogram is formed for each block of cells. There is an overlap of half the block size between consecutive blocks to increase accuracy. Finally, the histograms from the individual cells are normalized and concatenated to form the HOG vector. This whole operation of forming histograms and concatenating them is shown in the bottom row of

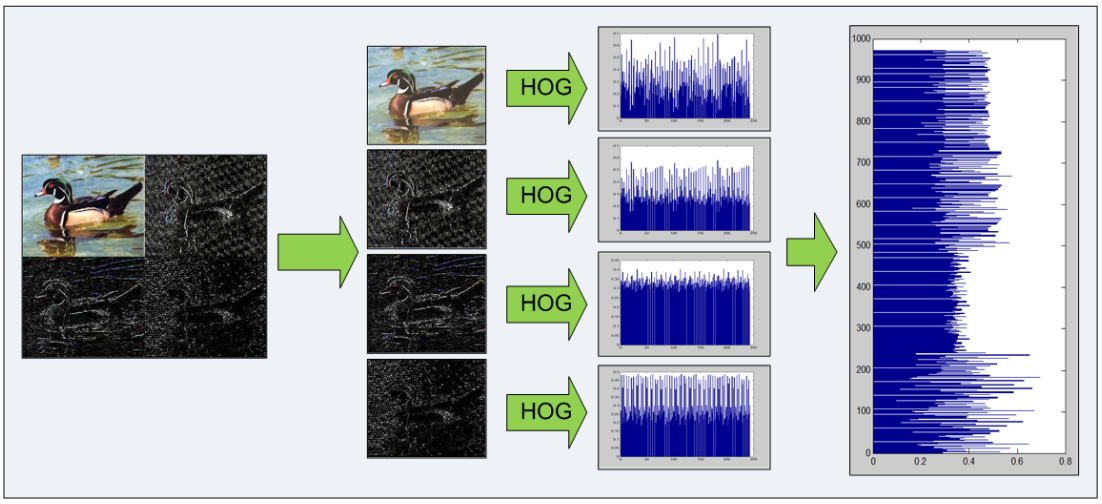


Figure 4.5 A color Haar wavelet transformed image, its four quadrant color images, their HOG descriptors, and their concatenation, the HaarHOG descriptor.

Figure 4.3.

Figure 4.4 shows how the same process is implemented for a color image. In particular, the color image shown on the left is split into three components and the gradient magnitudes and directions at each pixel are calculated for each of the three components separately. The gradient magnitude images for the three components are shown in the second column of the figure. Then each image window is divided into blocks of cells as described above and a HOG descriptor is calculated from each component image as shown in the third and fourth columns of Figure 4.4. Finally, these three HOG vectors are concatenated to get the color HOG descriptor.

4.3 An Innovative HaarHOG Descriptor

The motivation for the next proposed descriptor, the HaarHOG descriptor, is based on enhancing useful and important local high-frequency features before extracting shape for object and scene image classification. Towards that end, the Haar wavelet transform of an image is first computed. Then the HOG of the Haar wavelet transformed image is derived for encoding both shape and local features.

In particular, the Haar wavelet transform is first applied to an image to extract local information by enhancing local contrast. This process divides the image into four sub-images. One of these sub-images contains low frequency information and the other three contain the high frequency information in different orientations. Each of these sub-images are one-fourth the size of the original image.

To generate the new HaarHOG descriptor, the HOG is next calculated from the Haar wavelet transformed image, not as a whole but as a collection of four sub-images. Specifically, four HOG descriptors are computed from the four quadrants of a Haar wavelet transformed image and then concatenated to get the HaarHOG descriptor..

The process described above is applicable only to grayscale images. Since color images contain more discriminatory information than grayscale images, this information can be incorporated into the descriptor by calculating a HaarHOG vector from each color component image, and then concatenating the three vectors. This method is explained in Figure 4.5. Specifically, the four quadrants of the color Haar transformed image on the left of the figure undergo the HOG operation, and their vectors are concatenated to form the innovative color HaarHOG descriptor. The color image may be converted to the HSV, the YCbCr, the oRGB or any other color space from the RGB color space to obtain the color HaarHOG descriptor in the desired color space as the end result.

4.4 Experiments

This section first introduces the datasets used for testing this new image descriptor and then does a comparative assessment of the classification performance of the HaarHOG descriptor, the HOG descriptor, and some other popular image descriptors.

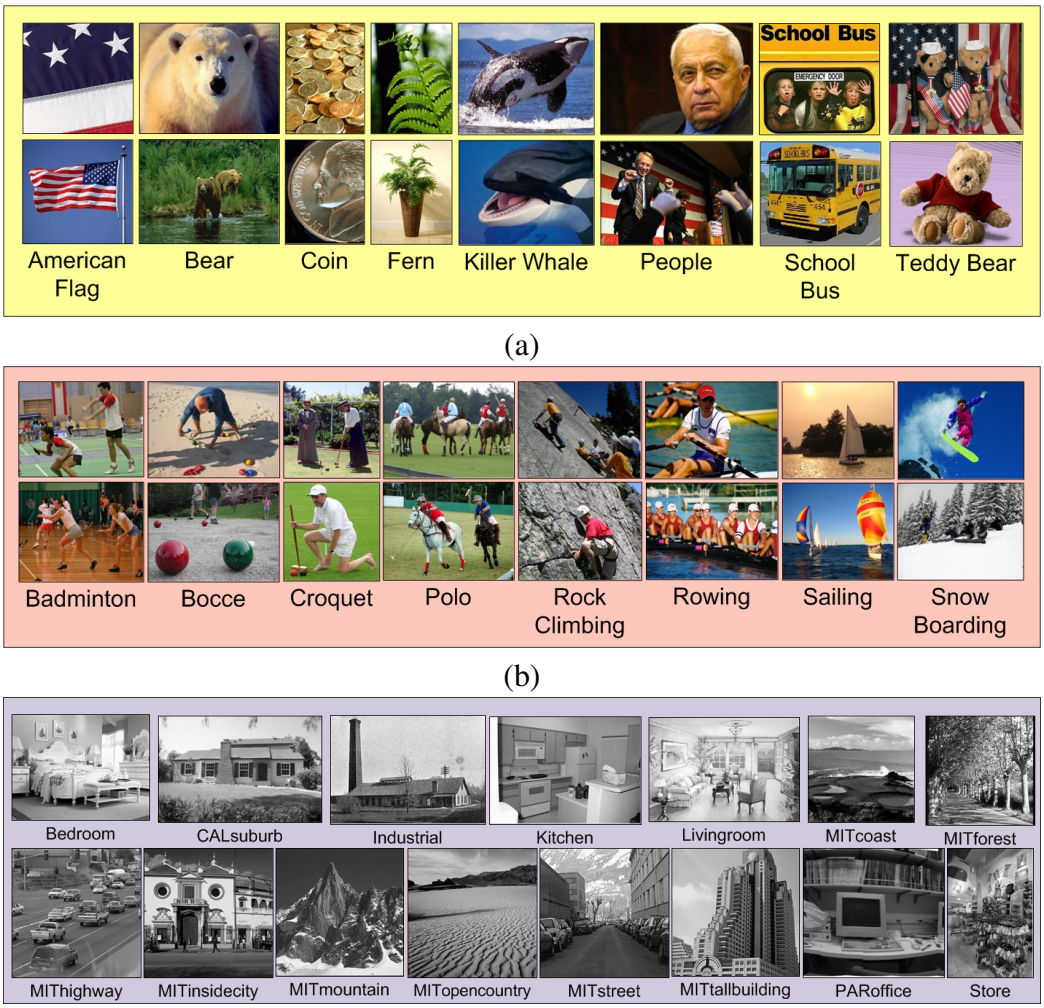


Figure 4.6 Some sample images from (a) the Caltech 256 dataset, (b) the UIUC Sports Event dataset, and (c) the Fifteen Scene Categories dataset. Please note that only a few classes from the Caltech 256 dataset are shown here.

4.4.1 Datasets Used

This section briefly introduces the four publicly available and widely used image datasets used for assessing the classification performance of the proposed descriptor.

The Caltech 256 Dataset

The Caltech 256 dataset (Griffin et al. 2007) holds 30,607 images divided into 256 object categories and a clutter class. The images have high intra-class variability and high object

location variability (Griffin et al. 2007). Each category contains a minimum of 80 images and a maximum of 827 images. The mean number of images per category is 119. The images represent a diverse set of lighting conditions, poses, backgrounds, and sizes (Griffin et al. 2007). Images are in color, in JPEG format with only a small percentage in grayscale. The average size of each image is 351×351 pixels. Some sample images from this dataset are shown in Figure 4.6(a).

The UIUC Sports Event Dataset

The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains 1,574 images from eight sports event categories: 250 rowing, 200 badminton, 182 polo, 137 bocce, 190 snowboarding, 236 croquet, 190 sailing, and 194 rock climbing. The mean image size in this dataset is 966×1156 pixels. These images contain both indoor and outdoor scenes where the foreground contains elements that define the category. The background is often cluttered and is similar across different categories like rowing and sailing, or croquet and polo. Some sample images from this dataset are shown in Figure 4.6(b).

The MIT Scene Dataset

The MIT Scene dataset, also known as the OT Scenes dataset (Oliva and Torralba 2001) has 2,688 images divided into eight categories. A detailed description of this dataset is provided in Section 3.4.1. Figure 3.4(a) shows some images from this dataset.

The Fifteen Scene Categories Dataset

The Fifteen Scene Categories dataset (Lazebnik et al. 2006) is composed of 15 scene categories: thirteen were provided by (Fei-Fei and Perona 2005), eight of which were originally collected by (Oliva and Torralba 2001) as the MIT Scene dataset, and two were collected by (Lazebnik et al. 2006). Each category has 200 to 400 images, most of which are grayscale.

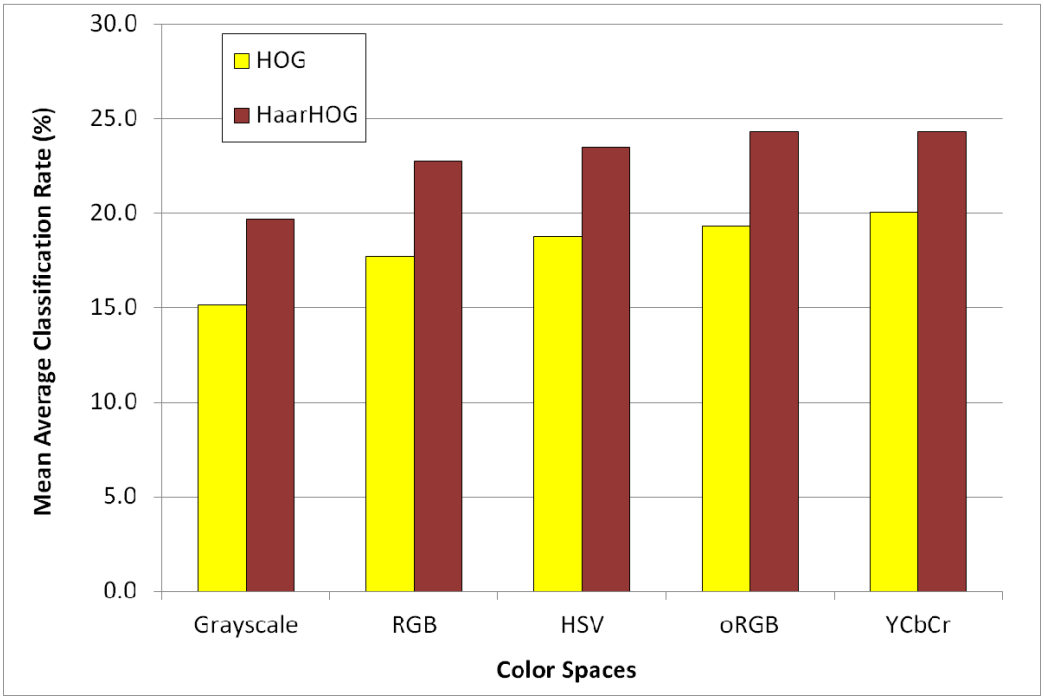


Figure 4.7 The mean average classification performance of the HOG and proposed HaarHOG descriptors in the grayscale, RGB, HSV, oRGB, and YCbCr color spaces using the SVM classifier on the Caltech 256 dataset.

Figure 4.6(c) shows one image each from the fifteen classes of this dataset.

4.4.2 Comparative Assessment of the HOG and HaarHOG Descriptors on the Different Datasets

In this section, a comparative assessment of the HOG and the proposed HaarHOG descriptors is made in four different color spaces – RGB, HSV, oRGB, and YCbCr color spaces, as well as in grayscale, using the four datasets described earlier to evaluate classification performance. Towards that end, first the descriptors are derived from each image in the different color spaces. Note that the large-scale images are resized in such a way that their largest dimension does not exceed 400 pixels. Each input image is converted into grayscale as well as transformed into images in the four color spaces, and the HOG and the HaarHOG descriptors are then computed from these images. For evaluating the relative classification

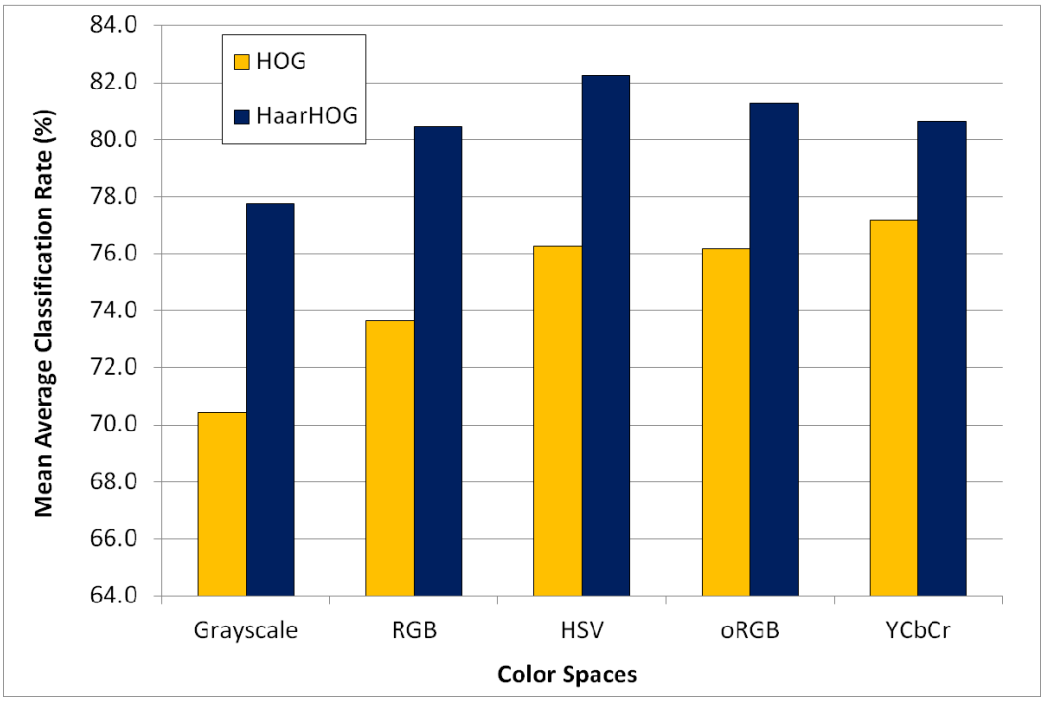


Figure 4.8 The mean average classification performance of the HOG and proposed HaarHOG descriptors in the grayscale, RGB, HSV, oRGB, and YCbCr color spaces using the SVM classifier on the UIUC Sports Event dataset.

performances of the HOG and HaarHOG descriptors, a Support Vector Machine (SVM) classifier with a linear kernel (Vapnik 1995; Vedaldi and Fulkerson 2010) is used.

From the Caltech 256 dataset, 50 images per class are used for training and 25 images per class are used for testing. The experiment is done for five random splits of the data with no overlap between training and testing sets of the same split. As can be seen in Figure 4.7, the HaarHOG significantly outperforms the HOG in all four color spaces as well as in grayscale. The horizontal axis shows the proposed descriptors in four different color spaces and in grayscale, and the vertical axis denotes the mean average classification performance, which is the percentage of correctly classified images averaged across all the 256 classes and five runs of experiments.

For the UIUC Sports Event dataset, 70 images are used from each class for training and 60 from each class for testing. Figure 4.8 shows the mean average classification performance obtained over five random splits of the data. Here also, the HaarHOG outperforms

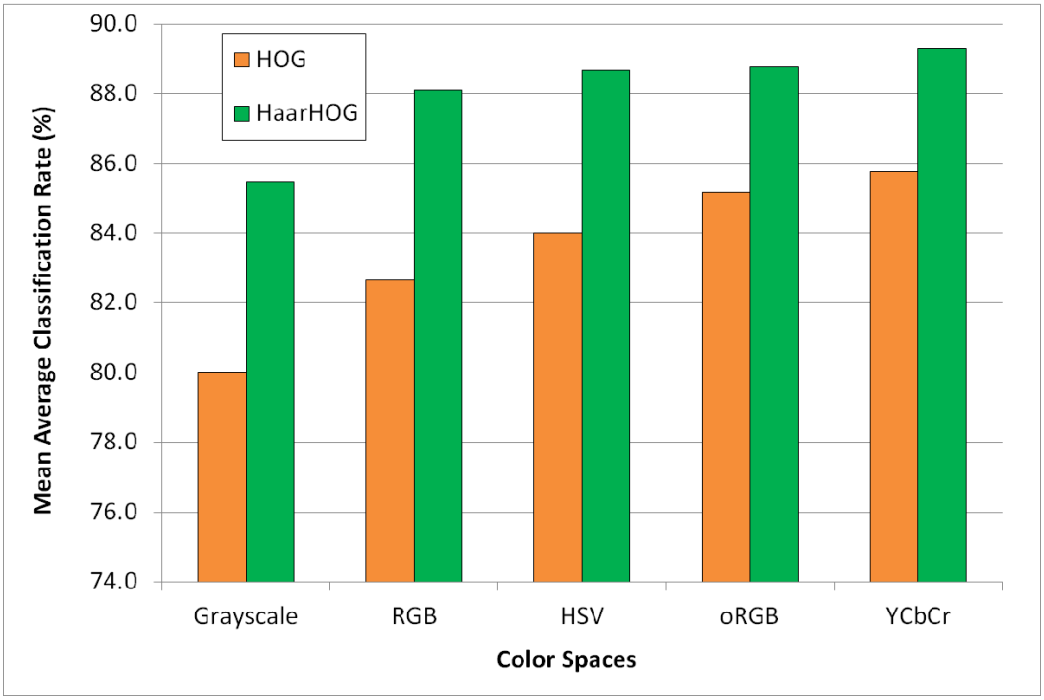


Figure 4.9 The mean average classification performance of the HOG and proposed HaarHOG descriptors in the grayscale, RGB, HSV, oRGB, and YCbCr color spaces using the SVM classifier on the MIT Scene dataset.

the HOG by a big margin that varies from about 3% to over 7%. Indeed, on this dataset the HaarHOG not only outperforms the HOG, but also provides a decent classification performance by itself.

From both the MIT Scene dataset and the Fifteen Scene Categories dataset five random splits of 100 images per class are used for training, and the rest of the images for testing. Again, the HaarHOG produces decent classification performance on its own apart from beating the HOG by a fair margin. Figure 4.9 displays these results on the MIT Scene dataset. Again, the horizontal axis shows the different descriptors in the four color spaces and in grayscale, and the vertical axis the mean average classification performance. The highest classification rate for this dataset is as high as 89.3% for the HaarHOG descriptor in the YCbCr color space which is a very respectable value for a dataset of this size and complexity. On the Fifteen Scene Categories dataset, experiments are conducted only in grayscale. The overall success rate for HOG on this dataset is 60.9% and for HaarHOG it

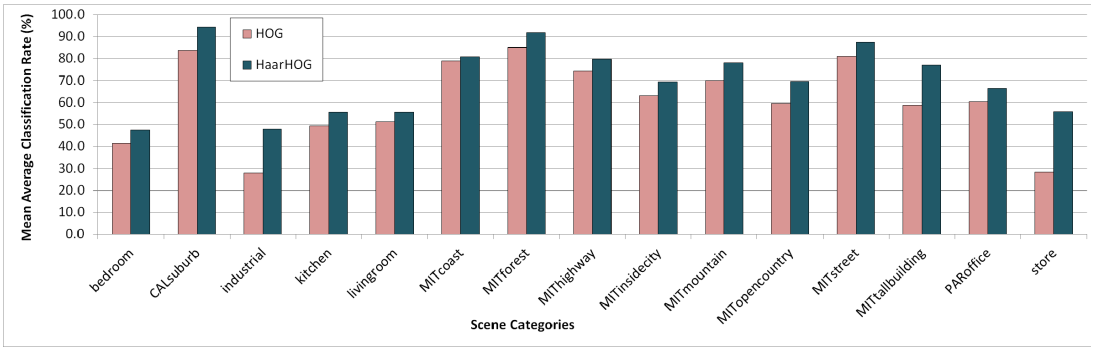


Figure 4.10 The comparative mean average classification performance of the HOG and HaarHOG descriptors on the 15 categories of the Fifteen Scene Categories dataset.

is 70.5%. In Figure 4.10 the category wise classification rates of the grayscale HOG and HaarHOG descriptors for all 15 categories of this dataset are displayed. Here, the horizontal axis reveals the fifteen scene categories, and the vertical axis displays the mean average classification performance. The HaarHOG here is shown to better the HOG classification performance in each scene category.

The classification performance of the proposed HaarHOG descriptor is also compared with some popular image classification techniques used by other researchers. The detailed comparison is shown in Table 4.1. It should be noted that the results of other researchers are reported directly from their published work.

4.5 Summary

In this chapter, a new image descriptor based on shape and local features that improves upon the popular HOG descriptor has been presented for object and scene image classification. First the new HaarHOG descriptor has been presented for a grayscale image. Then this definition has been extended for color images.

Also the HaarHOG descriptor has been comparatively assessed in grayscale and four different color spaces — the RGB, the HSV, the YCbCr, and the oRGB — for image classification performance. Experimental results using four datasets show that the proposed

Table 4.1 Comparison of the Classification Performance (%) of the Proposed HaarHOG Descriptor with Other Popular Methods on the UIUC Sports Event and MIT Scene Datasets

Descriptor		**UIUC Performance (%)**	**MIT Performance (%)**
SIFT+GGM	(Li and Fei-Fei 2007)	73.4	-
OB	(Li et al. 2010)	76.3	-
CA-TM	(Niu et al. 2012)	78.0	-
LBP		-	77.9
CGLF	(Banerji et al. 2011)	-	80.0
SE	(Oliva and Torralba 2001)	-	83.7
CGLF+PHOG	(Banerji et al. 2011)	-	84.3
C4CC	(Bosch et al. 2006)	-	86.7
HOG		76.3	85.8
HaarHOG	**(proposed)**	**82.2**	**89.3**

new HaarHOG descriptor not only achieves significantly better image classification performance than the conventional HOG descriptor, but can also beat other popular descriptors, such as the Scale Invariant Feature Transform (SIFT), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank (OB), Context Aware Topic Model (CA-TM), as well as LBP.

CHAPTER 5

THE NEW 3D-LBP, 3DLH AND 3DLH-FUSION DESCRIPTORS

As discussed in Chapter 3, the Local Binary Patterns (LBP) method describes the texture information of a grayscale image but is not effective in representing color. Considering the fact that a color image contains much more discriminative information than a grayscale image, and the performance of LBP alone on the MIT Scene dataset leaves scope for improvement, this chapter introduces a novel Three-dimensional Local Binary Patterns (3D-LBP) descriptor that attempts to encode both color and texture information from a color image. Further, the HaarHOG descriptor is fused with the 3D-LBP to produce the new 3DLH and 3DLH-fusion vectors that perform well for classification on different scene datasets.

5.1 A New Three-Dimensional Local Binary Patterns (3D-LBP) Descriptor

As discussed in Section 3.1, the Local Binary Patterns (LBP) method describes the texture information of a grayscale image by comparing each pixel with its neighbors (Ojala et al. 1994, 1996, 2002). LBP, however, does not encode color information, which is an effective cue for pattern recognition such as object and scene image classification (Banerji et al. 2011; Liu 2008, 2006). The motivation for this new three dimensional LBP descriptor, or 3D-LBP descriptor, rests on the extension of the conventional LBP method to incorporate the color cue when encoding a color image. Specifically, given a color image, the 3D-LBP descriptor generates three new color images by applying three perpendicular LBP encoding schemes. Figure 5.1 shows a color image, the three perpendicular LBP encoding schemes, and the three encoded color images generated by the proposed 3D-LBP descriptor. The first LBP encoding scheme applies a 3×3 neighborhood, which is shown in pink color in the top row of the second column, to encode the red, green, and blue component images, respectively. The encoded three images then form a new color image that is displayed as

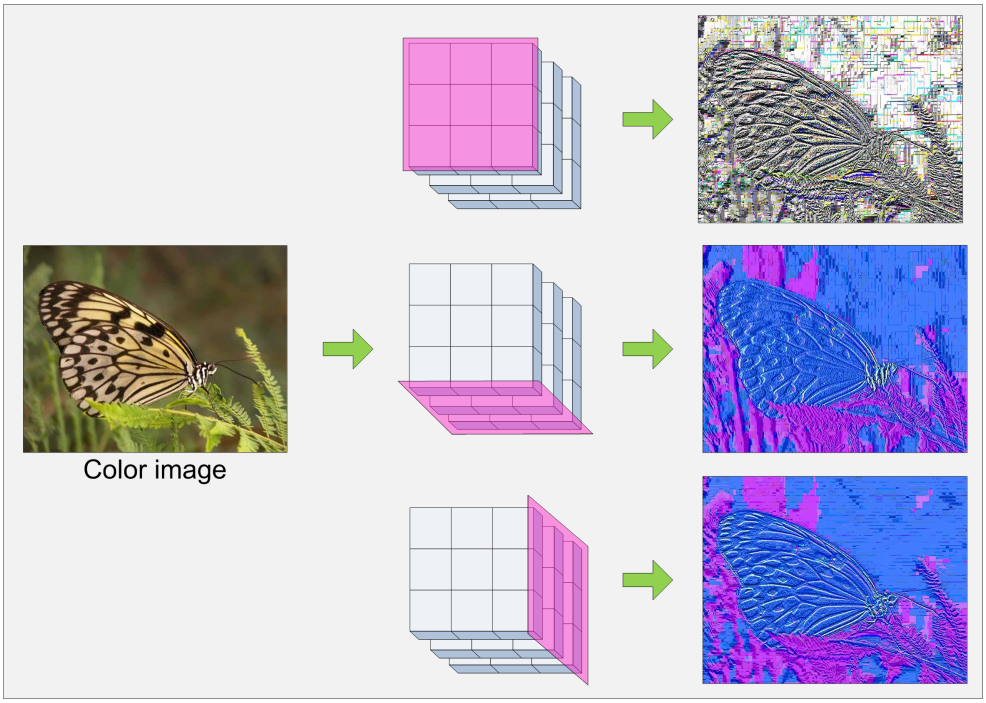


Figure 5.1 A color image taken from the butterfly category of the Caltech 256 dataset, the three perpendicular LBP encoding schemes, and the three encoded color images generated by the 3D-LBP descriptor.

the top image in the last column in Figure 5.1. The outer pixels are discarded on all sides after performing the LBP operation and hence this image is smaller than the original image by one pixel on all sides. The second LBP encoding scheme utilizes a 3×3 neighborhood shown in pink color in the middle row of the second column to encode the rows across the red, green, and blue component images, and the encoded three images form a new color image that is shown as the middle image in the last column in Figure 5.1. The third LBP encoding scheme uses a 3×3 neighborhood shown in pink color in the bottom row of the second column to encode the columns across the red, green, and blue component images, and the encoded three images form a new color image that is displayed as the bottom image in the last column in Figure 5.1. Normally, after performing an LBP operation, the outer pixels are discarded. However, since the number of color planes is just three, here the top and bottom planes cannot simply be discarded after performing the new LBP operations as

shown in the second and third rows of Figure 5.1. To solve this problem, the existing planes are replicated in a manner that puts an extra plane on either side of the three existing planes without copying a plane next to itself. For example, if the image is RGB, the new five-plane matrix will be BRGBR. After the 3D-LBP operation is done, these two new planes, i.e. the first and fifth planes of the five-plane image, are discarded to create a three plane image again. The 3D-LBP descriptor thus encodes the color and texture information to generate three new color images as shown in the last column in Figure 5.1. The histograms are taken from each color plane of these three images and concatenated to form the 3D-LBP feature vector which is independent of the image size. Hence, for a color image, the size of the 3D-LBP vector is 2304, which is $256 \times 3 \times 3$.

5.2 The 3DLH Descriptor

To create a descriptor that encodes color, texture, shape and local features, the new 3D-LBP descriptor and the HaarHOG feature vector introduced in Chapter 4 are fused. To implement this, the two feature vectors are subjected to dimensionality reduction by PCA and the most expressive features thus obtained are concatenated to form the 3DLH feature vector. In particular, first a color image is subjected to the 3D-LBP operation which produces three new color images. The histograms of each of the three planes of these three color images are concatenated to form the 3D-LBP feature vector. Next, each color component image of the original color image is subjected to Haar transformation to produce three images that are divided into four quadrants each. Then the HOG feature vector is computed from each of the four quadrants of each image and concatenated to form the HaarHOG descriptor. Please note that the HOG operation used in this chapter divides the image into 3×3 blocks and uses nine histogram bins for the orientations. Finally, the 3D-LBP feature vector and HaarHOG feature vector are both subjected to dimensionality reduction by PCA and the results are concatenated to form the 3DLH descriptor. Figure 5.2 shows a color image, its 3D-LBP images, its 3D-LBP feature vector, its HaarHOG feature vector, and the novel

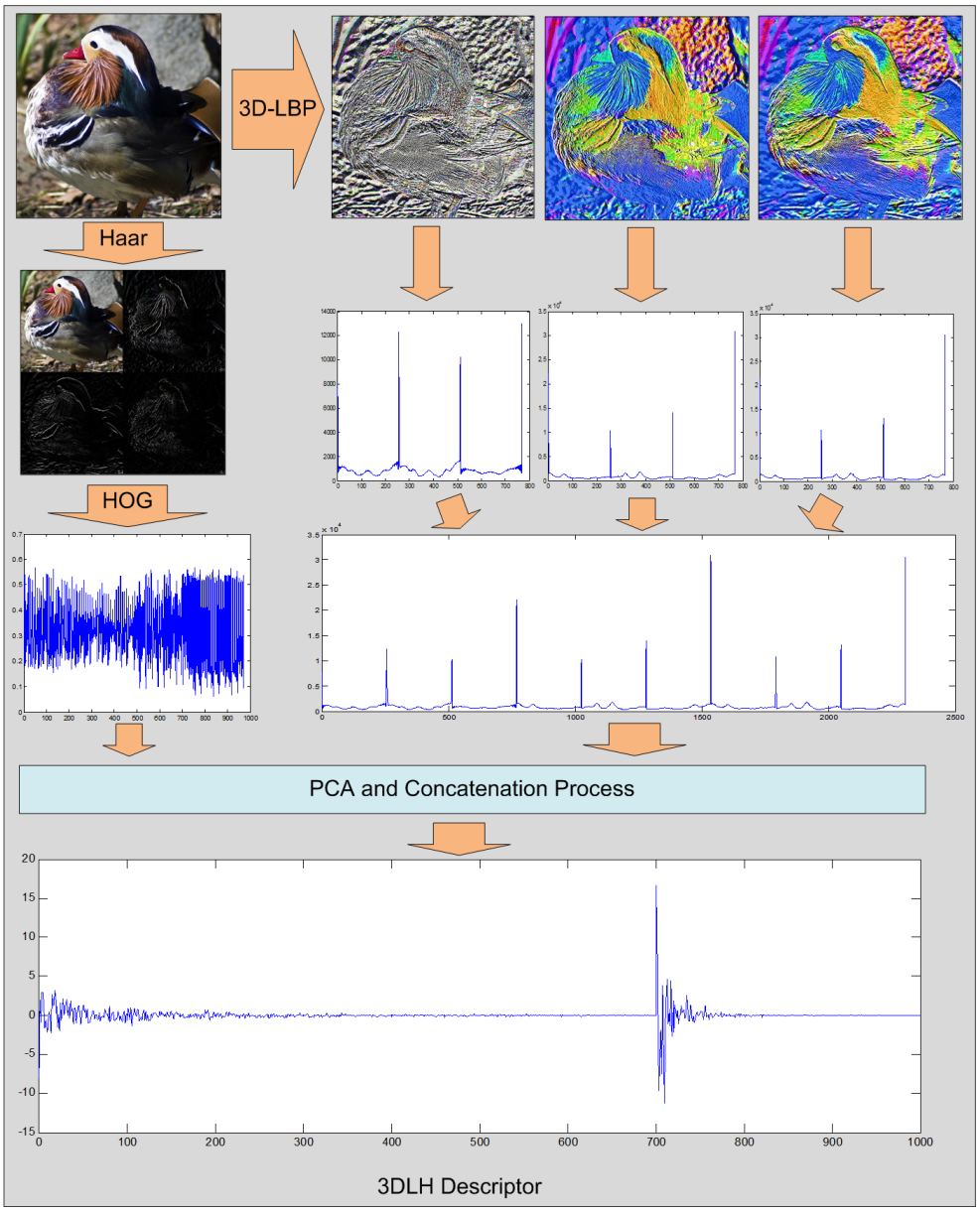


Figure 5.2 A color image, its three 3D-LBP color images, its Haar transformed image, the HaarHOG and 3D-LBP histogram descriptors, the PCA process and the concatenated 3DLH descriptor.

3DLH feature vector. In particular, the top left image of Figure 5.2 shows a color image of a Mandarin duck and the leftmost column shows the original color image undergoing Haar wavelet transformation to generate a new color image shown in the second row, from which the HaarHOG descriptor shown in row 3 is computed. On the right side of the original im-

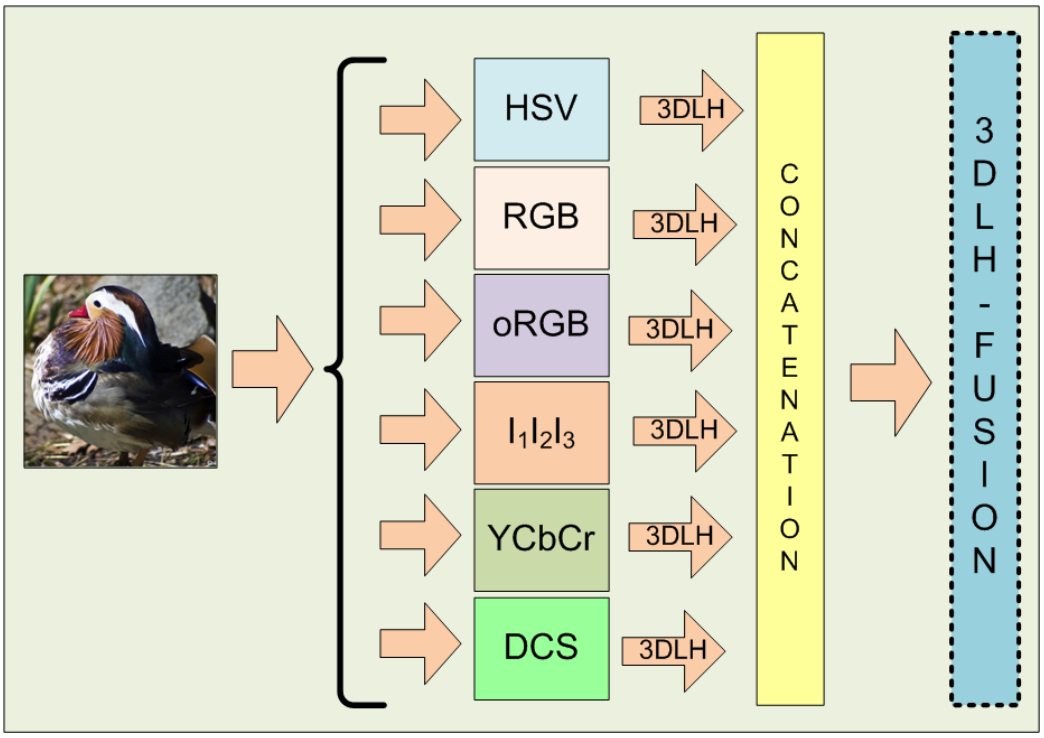


Figure 5.3 A schematic diagram showing a color image, and the process of computing its 3DLH-fusion descriptor by concatenating its 3DLH descriptors in six color spaces.

age on the top row, the three new color images formed from the original color image by the 3D-LBP operation are shown. Each of these three images are used for extracting their histograms from all color planes and finally these histograms are concatenated to form the 3D-LBP feature vector. This vector is shown on the right side of the third row of the figure. Finally, the HaarHOG and 3D-LBP descriptors undergo dimensionality reduction by PCA and concatenation to form the 3DLH vector shown in the last row of the figure.

5.3 A Novel 3DLH-fusion Descriptor

Color contains a significant amount of discriminatory information for object and scene image classification (Liu and Mago 2012; Liu 2011; Liu and Yang 2009; Liu 2008, 2007, 2006, 2004) and various color spaces have different properties that are useful to encode all the information present in images. Further, it has been shown by (Verma et al. 2010; Banerji

et al. 2011) that different color spaces are not fully redundant in their information content. To further incorporate color information to the proposed 3DLH descriptor, an innovative 3DLH-fusion descriptor is introduced that fuses the 3DLH descriptors in six different color spaces, where the color spaces are the RGB, oRGB, HSV, YCbCr, $I_1I_2I_3$ and DCS color spaces. Detailed descriptions of these color spaces are available in Section 2.1. Specifically, the original RGB color image is transformed to each of the other five color spaces, and the 3DLH descriptor is computed for each color space as described in Section 5.2. Finally, the descriptors from all six color spaces are concatenated to form the 3DLH-fusion feature vector. Figure 5.3 shows a color image, and the formation of the 3DLH-fusion descriptor from it. In particular, it shows a color image on the left, then its conversion into different color spaces is shown schematically. Next, the formation of the 3DLH feature vector from each color space is shown, followed by concatenation and formation of the 3DLH-fusion descriptor shown in blue on the right.

5.4 Experiments

The proposed descriptors are tested for scene image classification using three challenging datasets, namely the Caltech 25 Scene dataset which is a small subset of the Caltech 256 dataset (Griffin et al. 2007), the UIUC Sports Event dataset (Li and Fei-Fei 2007), and the MIT Scene dataset (Oliva and Torralba 2001). Specifically, the 3DLH descriptor is first assessed in six different color spaces, and then the 3DLH-fusion descriptor is compared with other popular descriptors, such as combinations of Scale Invariant Feature Transform (SIFT) (Lowe 1999, 2004) with other descriptors (Li and Fei-Fei 2007; Bo et al. 2011), LBP (Ojala et al. 1994) and Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b) based descriptors (Banerji et al. 2011), Spatial Envelope (Oliva and Torralba 2001), Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), and other approaches such as Context Aware Topic Model (CA-TM) (Niu et al. 2012), and Object Bank (Li et al. 2010).

Figure 5.4 Some sample images from the Caltech Scene 25 dataset.

5.4.1 Datasets Used

This section briefly introduces the three publicly available image datasets. Two of these datasets are quite popular among researchers for evaluating the performance of scene image descriptors and classification methods. The remaining one is a subset of the very large and challenging Caltech 256 dataset.

The Caltech 25 Scene Dataset

The Caltech 256 dataset (Griffin et al. 2007) contains 30,607 images distributed among 256 object categories and a clutter class. The images are diverse, with high intra-class variability and large variations in object location, lighting and camera angles. The complete dataset is described in greater detail in Section 4.4.1. For testing the 3DLH descriptor, 25 scene image categories are selected from this dataset to form the Caltech 25 Scene dataset. This subset contains 110 Camel, 104 Canoe, 87 Duck, 83 Eiffel Tower, 101 Elk, 110 Fern, 100 Fireworks, 80 Golden Gate Bridge, 201 Grapes, 93 Hawksbill, 120 Ibis, 108 Iris, 111 Ketch, 91 Killer whale, 190 Leopards, 136 Lightning, 130 Minaret, 202 Mushroom, 103 Palm tree, 102 Rainbow, 95 Skyscraper, 105 Tennis court, 90 Tower Pisa, 95 Waterfall and 96 Zebra images. The images, while representing a diverse set of lighting conditions, backgrounds and sizes, are from categories that can be called scenes. As Figure 5.4 shows,

some classes like Minaret and Tower-Pisa have a very similar foreground, and others like Lightning and Fireworks have a similar background, making classification very challenging. Almost all of these images are in color JPEG format with only a few in grayscale.

The UIUC Sports Event Dataset

The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains eight sports event categories. This dataset has been described in detail in Section 4.4.1. Some sample images are displayed in Figure 4.6(b).

The MIT Scene Dataset

The MIT Scene dataset (Oliva and Torralba 2001) has 2,688 color jpeg images divided into eight scene categories. A more elaborate description of this dataset is given in Section 3.4.1. Some sample images from this dataset are shown in Figure 3.4(a).

5.4.2 Comparative Assessment of the 3DLH Descriptor in Different Color Spaces

Now the 3DLH descriptor is evaluated in six different color spaces — the RGB, oRGB, HSV, YCbCr, DCS and $I_1I_2I_3$ color spaces — for content-based image classification performance using the three datasets mentioned above. The larger images are resized to reduce their longer dimension to 400 pixels. To extract the 3DLH descriptor from each image, the 3D-LBP descriptor is first computed to produce three new color images. Then the HaarHOG descriptor of the original color image is calculated. The 3DLH feature vector is computed by taking the most expressive features from both the 3D-LBP and HaarHOG feature vectors using PCA and then concatenating them. Each image is converted to the six color spaces and then processed in the same way to construct the six different color 3DLH descriptors. Next, PCA is applied to reduce the dimensionality of the 3DLH descriptors, and the features thus extracted are further processed by EFM to obtain the features that

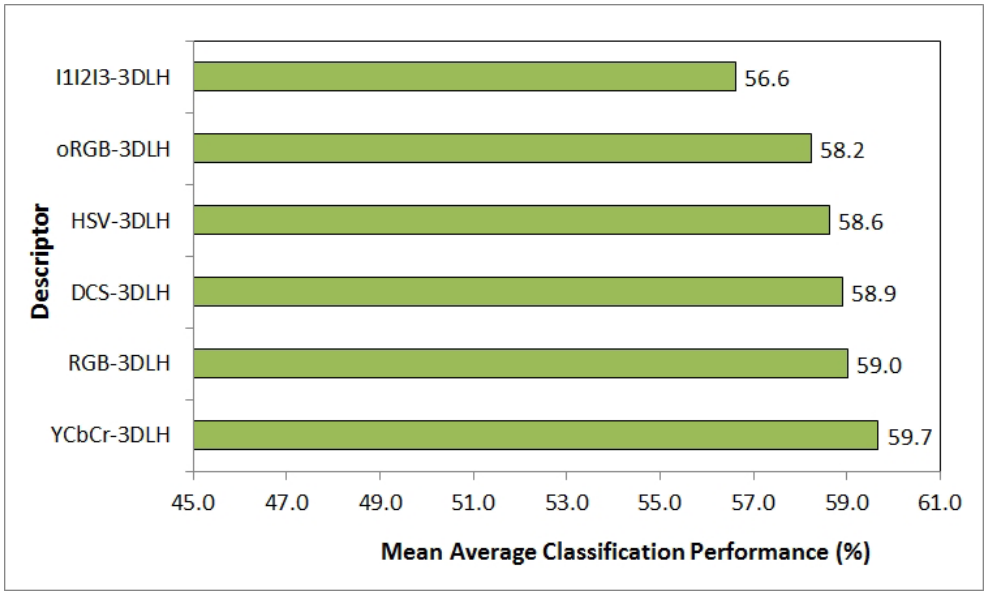


Figure 5.5 The mean average classification performance of the proposed 3DLH descriptor in the $I_1I_2I_3$, oRGB, HSV, DCS, RGB, and YCbCr color spaces using the EFM-NN classifier on the Caltech 25 Scene dataset.

are most discriminatory for classification. Finally a nearest neighbor classifier is used for image classification.

For the Caltech 25 Scene dataset, experiments are conducted for the 3DLH descriptors from six different color spaces. 50 images from each class are used for training and 25 images for testing. Figure 5.5 shows the detailed performance of the 3DLH descriptor coupled with the EFM-NN classification technique on this dataset. The horizontal axis indicates the average classification performance, which is the percentage of correctly classified images averaged across the 25 classes and the five random runs of the experiments, and the vertical axis shows the six different 3DLH descriptors in the six color spaces. Among the different 3DLH descriptors, the best recognition rate obtained is 59.7% for the 3DLH descriptor in the YCbCr color space. The classification rates for the 3DLH descriptor in the RGB, DCS, HSV, oRGB and $I_1I_2I_3$ color spaces are 59.0%, 58.9%, 58.6%, 58.2% and 56.6%, respectively.

In the case of the UIUC Sports Event dataset, the protocol defined in (Li and Fei-

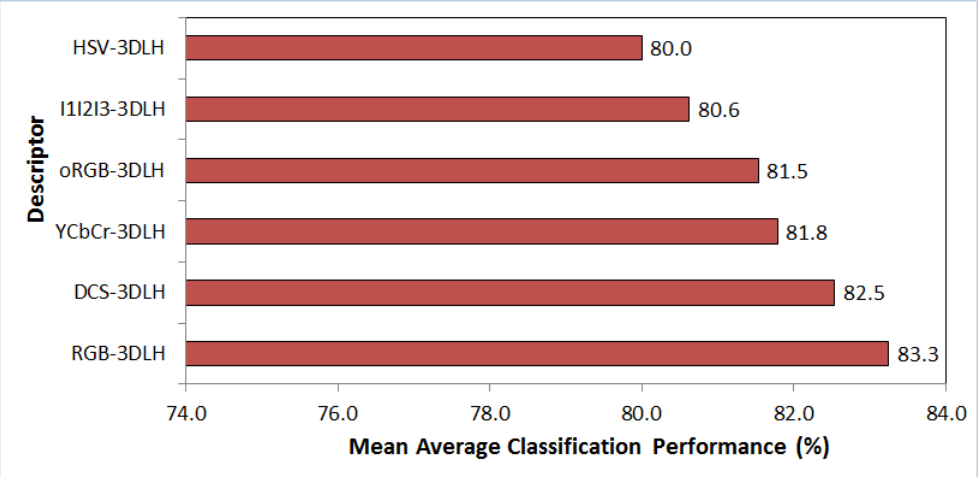


Figure 5.6 The mean average classification performance of the proposed 3DLH descriptor in the HSV, $I_1I_2I_3$, oRGB, YCbCr, DCS and RGB color spaces using the EFM-NN classifier on the UIUC Sports Event dataset.

Fei 2007) is used, which specifies that 70 images from each class are used for training and 60 images for testing. To reduce variability in performance, these experiments are repeated five times using random splits of the data, with no overlap between the training and the testing sets of the same split. Figure 5.6 reveals that the RGB-3DLH is the best descriptor with 83.3% average classification performance followed in order by the 3DLH descriptors in the DCS, YCbCr, oRGB, $I_1I_2I_3$, and HSV color spaces with 82.5%, 81.8%, 81.5%, 80.6% and 80.0% success rates, respectively. Again the horizontal axis indicates the average classification performance and the vertical axis the 3DLH descriptors in the six color spaces.

For the MIT Scene dataset, 250 images are used from each class for training and the rest of the images are used for testing. Here too, all experiments are performed for five random splits of the data. The bar chart in Figure 5.7 displays that the 3DLH descriptor in the YCbCr color space performs the best with 88.4% average classification rate. The 3DLH descriptors in the HSV, oRGB, DCS, $I_1I_2I_3$ and RGB color spaces correctly classify on an average 88.3%, 88.2%, 87.9%, 85.8% and 85.3% of the images respectively. Here also, the horizontal axis shows the average classification performance and the vertical axis

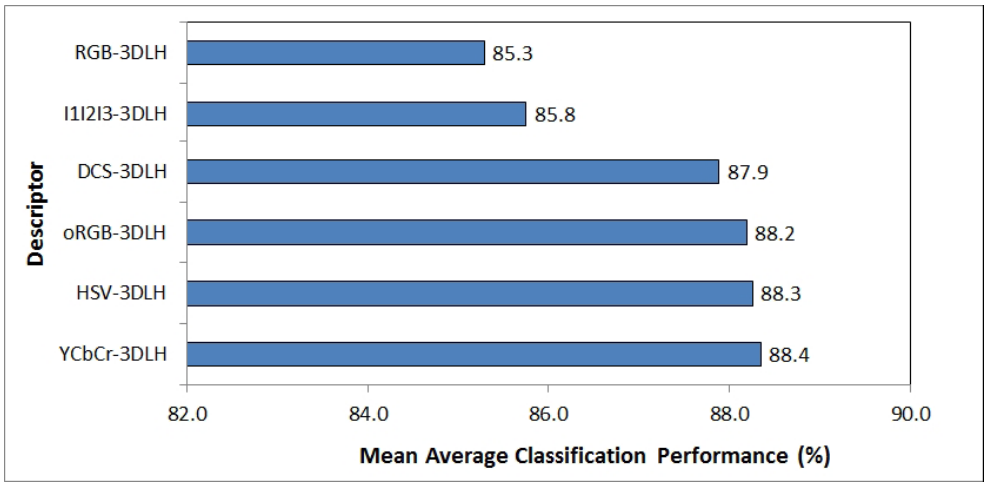


Figure 5.7 The mean average classification performance of the proposed 3DLH descriptor in the RGB, $I_1I_2I_3$, DCS, oRGB, HSV, and YCbCr color spaces using the EFM-NN classifier on the MIT Scene dataset.

shows the different 3DLH descriptors.

5.4.3 Comparative Assessment of the 3DLH-fusion Descriptor and Some Popular State-of-the-art Image Descriptors

In this section the performance of the proposed 3DLH-fusion descriptor on the three datasets described in Section 5.4.1 is evaluated. First the proposed 3DLH-fusion descriptor is compared with the 3D-LBP-fusion and HaarHOG-fusion descriptors to show the improvement obtained by combining the descriptors. Then the 3DLH-fusion descriptor is compared with some other popular state-of-the-art descriptors using the image classification performance reported in the published papers.

To justify combining the 3D-LBP and the HaarHOG descriptors to form the 3DLH vector, the 3DLH-fusion descriptor is next compared with the 3D-LBP-fusion and the HaarHOG-fusion descriptors. The 3D-LBP-fusion descriptor is formed by concatenating the 3D-LBP descriptors from different color spaces to form one single feature vector. Similarly, the HaarHOG-fusion feature vector is formed by concatenating the HaarHOG descriptors from individual color spaces. Both descriptors are subjected to EFM-NN clas-

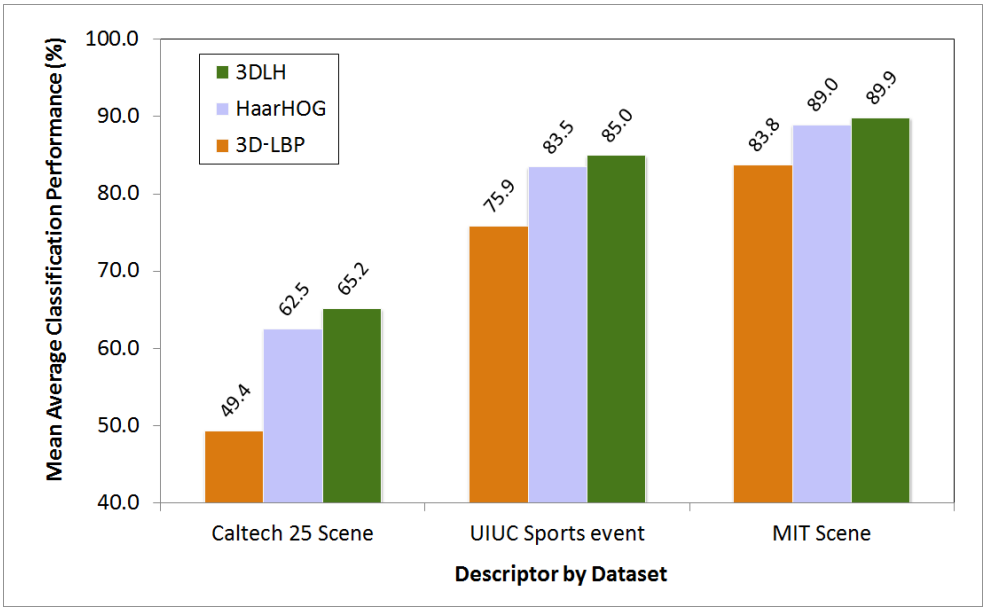


Figure 5.8 The comparative mean average classification performance of the 3D-LBP-fusion, HaarHOG-fusion and 3DLH-fusion descriptors on the Caltech 25 Scene, UIUC Sports Event and MIT Scene datasets.

sification for a fair comparison with the 3DLH-fusion descriptor.

Figure 5.8 shows that the 3DLH-fusion descriptor has an image classification performance better than both the 3D-LBP-fusion and the HaarHOG-fusion descriptors on all the three datasets. Note that the horizontal axis of this graph lists the three descriptors and the three datasets while the vertical axis shows the average classification performance as a percentage. In particular, on the Caltech 25 Scene dataset, the 3D-LBP-fusion, the HaarHOG-fusion and the 3DLH-fusion descriptors achieve the average classification rate of 49.4%, 62.5% and 65.2% respectively. On the UIUC Sports Event dataset, the 3DLH-fusion yields 85.0% classification rate, compared to the 3D-LBP-fusion descriptor with the average classification rate of 75.9% and to the HaarHOG-fusion descriptor with the average classification rate of 83.5%, respectively. On the MIT Scene dataset, the average classification rates for the 3D-LBP-fusion, the HaarHOG-fusion and the 3DLH-fusion descriptors are 83.8%, 89.0% and 89.9% respectively.

Using this UIUC Sports Event dataset, the 3DLH-fusion descriptor is further com-

Table 5.1 Comparison of the Classification Performance (%) of the 3DLH-fusion Descriptor with other Popular Methods on the UIUC Sports Event Dataset

Descriptor		Performance (%)
SIFT+GGM	(Li and Fei-Fei 2007)	73.4
OB	(Li et al. 2010)	76.3
CA-TM	(Niu et al. 2012)	78.0
SIFT+SC	(Bo et al. 2011)	82.7
3DLH-fusion	**(proposed)**	**85.0**

pared with some popular state-of-the-art descriptors and methods, such as the Context Aware Topic Model (CA-TM) (Niu et al. 2012), the Object Bank approach (Li et al. 2010) and variations of the popular Scale Invariant Feature Transform (SIFT) (Lowe 2004) descriptor (Bo et al. 2011; Li and Fei-Fei 2007). Note that the performance reported here for the competing methods are from the published papers. Table 5.1 shows that the 3DLH-fusion descriptor achieves the best classification performance of 85.0% compared to SIFT+SC (Bo et al. 2011) with classification performance of 82.7% , to Context Aware Topic Model (CA-TM) (Niu et al. 2012) with classification performance of 78.0%, to Object Bank (OB) (Li et al. 2010) with classification performance of 76.3% and to the SIFT+GGM (Li and Fei-Fei 2007) method with classification performance of 73.4%.

On the MIT Scene dataset, two sets of experiments are performed with the 3DLH-fusion descriptor. First 250 images are randomly selected from each class for training and the rest of the images are used for testing. In this set of experiments, the proposed 3DLH-fusion descriptor yields an average success rate of 89.9%. In the next set of experiments 100 images per class are used for training and the leftover images for testing. Next, the proposed descriptor is compared with some popular state-of-the-art descriptors and classification methods such as the Spatial Envelope (Oliva and Torralba 2001), Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), Color Grayscale LBP Fusion (CGLF) (Banerji et al. 2011), LBP and Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b; Banerji et al. 2011). Here also, the results achieved by other researchers are

Table 5.2 Comparison of the Classification Performance (%) of the 3DLH-fusion Descriptor with other Popular Methods on the MIT Scene Dataset

	#train = 2000, #test = 688	
3DLH-fusion	**Proposed Descriptor**	**89.9**
CGLF+PHOG	(Banerji et al. 2011)	89.5
CGLF	(Banerji et al. 2011)	86.6
LBP		82.7
PHOG	(Banerji et al. 2011)	79.1
	#train = 800, #test = 1888	
3DLH-fusion	**Proposed Descriptor**	**87.0**
C4CC	(Bosch et al. 2006)	86.7
CGLF+PHOG	(Banerji et al. 2011)	84.3
SE	(Oliva and Torralba 2001)	83.7
CGLF	(Banerji et al. 2011)	80.0
LBP		77.9

reported directly from their published work. Table 5.2 shows that with 250 training images, the proposed 3DLH-fusion descriptor shows the best classification performance of 89.9% as compared to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 89.5%, to CGLF (Banerji et al. 2011) with a classification performance of 86.6%, to LBP with a classification performance of 82.7% and to PHOG (Bosch et al. 2007b; Banerji et al. 2011) with a classification performance of 79.1%. With 100 training images per class, once again the 3DLH-fusion descriptor generates the best classification performance of 87.0%, as compared to Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006) with a classification performance of 86.7%, to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 84.3%, to Spatial Envelope with a classification performance of 83.7%, to CGLF (Banerji et al. 2011) with a classification performance of 80.0% and to LBP with a classification performance of 77.9%.

5.5 Summary

This chapter has presented new image descriptors based on color, texture, shape, and wavelets for scene image classification. In particular, a new LBP-based color and texture feature extraction method (3D-LBP) has been proposed for images and combined with Haar wavelet features and HOG features to generate several new descriptors for color scene images. Results of the experiments using three challenging datasets show that the oRGB-3DLH, HSV-3DLH, DCS-3DLH and YCbCr-3DLH descriptors improve recognition performance over several other popular descriptors. The fusion of multiple color 3DLH descriptors (3DLH-fusion) shows an increase in the classification performance, which suggests that various color 3DLH descriptors are not completely redundant for image classification. Also, the fusion of 3D-LBP and HaarHOG descriptors improves classification performance which indicates that these two descriptors contain non-redundant information and, if fused more effectively, could yield an even higher classification performance.

CHAPTER 6

THE NOVEL H-DESCRIPTOR AND H-FUSION DESCRIPTOR

This chapter presents new image descriptors that integrate color, texture, shape, and wavelets for object and scene image classification. First, the three Dimensional Local Binary Patterns (3D-LBP) descriptor is used for encoding the color and texture information of a color image. Specifically, the 3D-LBP descriptor produces three new color images from the original color image. Second, the Haar wavelet transform is applied to the three new 3D-LBP color images and the original color image. The Histograms of Oriented Gradients (HOG) of these Haar wavelet transformed images are further calculated for encoding shape and local features. Third, a novel H-descriptor is proposed, which integrates the 3D-LBP and the HOG of its wavelet transform, to encode color, texture, shape, and local information for object and scene image classification. Finally, a new H-fusion descriptor is presented by fusing the Principal Component Analysis (PCA) features of the H-descriptors in the seven individual color spaces.

Experimental results using three datasets, the Caltech 256 object categories dataset, the UIUC Sports Event dataset, and the MIT Scene dataset, show that the proposed new image descriptors achieve better image classification performance than other popular image descriptors, such as the Scale Invariant Feature Transform (SIFT) (Lowe 1999, 2004), the Pyramid Histograms of visual Words (PHOW) (Bosch et al. 2007a), the Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b; Banerji et al. 2011), Spatial Envelope (Oliva and Torralba 2001), Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), Object Bank (Li et al. 2010), the Hierarchical Matching Pursuit (Bo et al. 2011), as well as LBP (Ojala et al. 1994).

6.1 A Novel H-Descriptor Based on Color, Texture, Shape, and Wavelets

The 3D-LBP descriptor introduced in Section 5.1 improves upon the conventional LBP method by means of encoding both color and texture information of a color image. The HaarHOG descriptor presented in Section 4.3 incorporates additional useful and important features for object and scene image classification, namely shape and local features. The 3D-LBP and the HaarHOG were fused in Section 5.2 to form the new 3DLH feature vector that outperformed both the 3D-LBP and HaarHOG descriptors for scene image classification. The motivation for the next descriptor, the H-descriptor, arises from the need for a better technique to integrate the texture, color, shape and local features from an image so that classification performance can be further improved. Towards that end, the H-descriptor is created which is the concatenation of the HaarHOG features of the original color image and the three color images produced by the 3D-LBP operation.

Specifically, first three new color images are generated from a color image using the 3D-LBP operation. Next, the Haar wavelet transform of the original color image and its new 3D-LBP color images are computed. Then, four HOG descriptors are computed from the four quadrants of a Haar wavelet transformed image and then concatenated to get the HOG descriptor of a Haar wavelet transformed image. Finally the HOG descriptors from the Haar wavelet transform of the component images of the color image and its 3D-LBP color images are integrated to form the H-descriptor, which encodes color, texture, shape, and local information for object and scene image classification. Please note that the HOG operation used in this chapter also divides the image into 3×3 blocks and uses nine histogram bins for the orientations.

In particular, for a color image, the 3D-LBP descriptor first generates three new color images. The Haar wavelet transform then produces twelve wavelet transformed images from the twelve color component images of the color image and its three 3D-LBP color images. The HOG process further generates four HOG descriptors corresponding to each of the Haar wavelet transformed images. The HOG descriptors from all the Haar

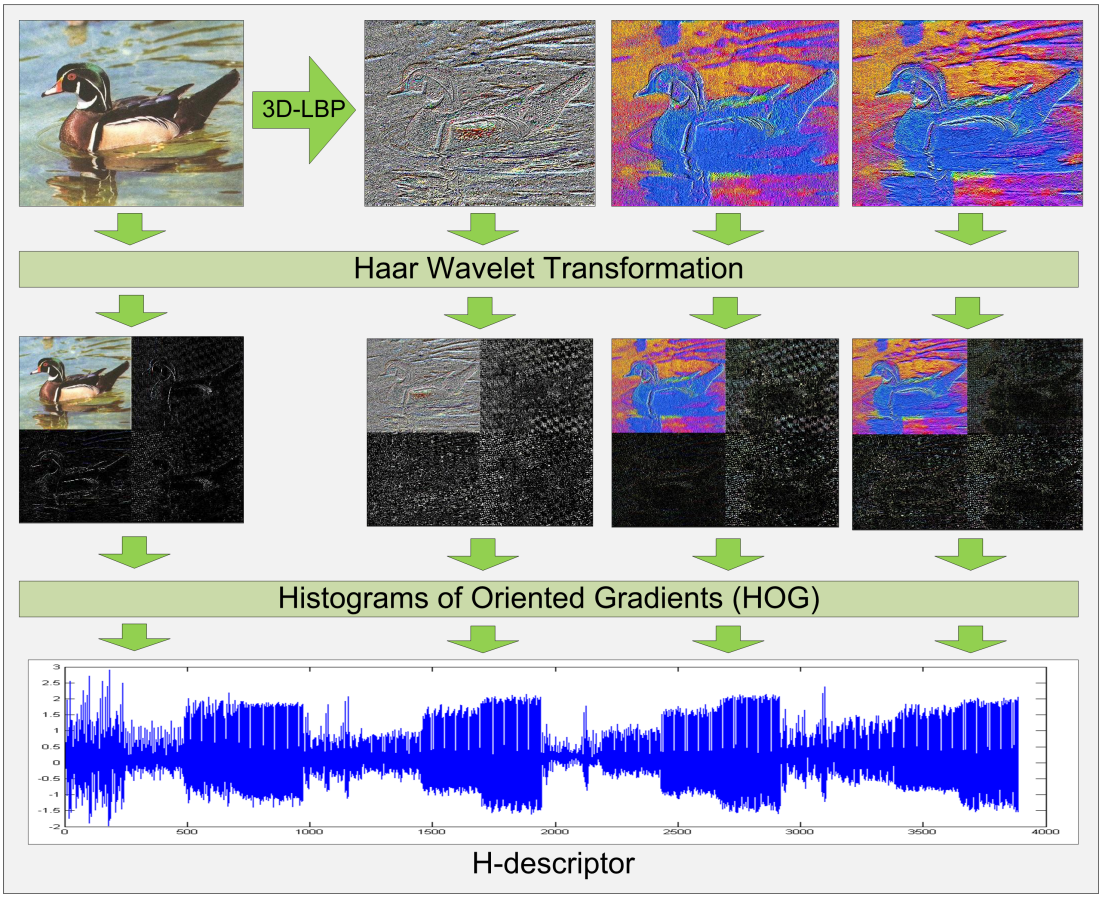


Figure 6.1 A color image taken from the duck category of the Caltech 256 dataset, its 3D-LBP color images, the Haar wavelet transforms of these color images, and the H-descriptor formed by the concatenation of the HOG descriptors of these Haar transform images.

wavelet transformed images are finally concatenated to form a new descriptor, the H-descriptor. The dimensionality of this descriptor is 3888 which is the product of the size of the grayscale HOG vector and the total number of quadrants from all the twelve component images of the four Haar transformed color images $(81 \times 4 \times 12)$. The time taken to compute the H-descriptor from an image is empirically seen to be directly proportional to the number of pixels in the image. For experiments done with a large number of images, the average feature extraction time is found to be 5.5 seconds per image on an Intel® Core™ i3-2120 3.30GHz CPU with 8 GB RAM. Figure 6.1 shows a color image, its 3D-LBP color images, the Haar wavelet transformed color images, and the H-descriptor derived from the

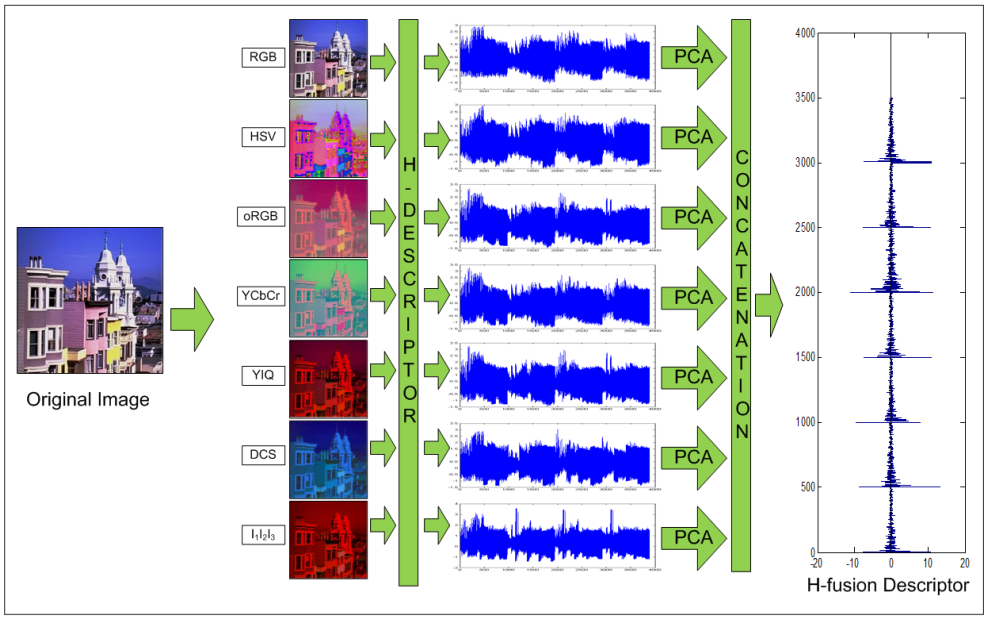


Figure 6.2 A color image taken from the inside-city category of the MIT Scene dataset, its corresponding color images in the seven color spaces, the H-descriptors of the color images, the PCA process, the concatenation process, and the H-fusion descriptor.

concatenation of the HOG descriptors of the Haar wavelet transformed color images.

6.1.1 An Innovative H-fusion Descriptor

Color provides a very important cue for pattern recognition in general and for object and scene image classification in particular (Liu and Mago 2012; Banerji et al. 2011; Liu 2011; Verma et al. 2010; Liu and Yang 2009; Liu 2008, 2007, 2006, 2004). To further incorporate color information, an H-fusion descriptor is introduced that fuses the most expressive features of the H-descriptors in seven different color spaces, where the most expressive features are extracted by means of principal component analysis (PCA) and the seven color spaces are the RGB, oRGB, HSV, YIQ, YCbCr, $I_1I_2I_3$, and DCS color spaces (Liu 2008). These color spaces have been described in detail in Section 2.1. PCA has been discussed in Section 2.2.

The proposed H-fusion descriptor is derived by first computing the H-descriptors

in the seven color spaces, namely RGB, oRGB, HSV, YIQ, YCbCr, $I_1I_2I_3$, and DCS, and then extracting the most expressive features of the H-descriptors using PCA, and finally concatenating these most expressive features from the seven color spaces. Figure 6.2 shows a color image, its corresponding color images in the seven color spaces, the H-descriptors of the color images, the PCA process, the concatenation process, and the H-fusion descriptor.

6.2 Experiments

The proposed descriptors are tested for object and scene image classification using three popular datasets, namely the Caltech 256 dataset (Griffin et al. 2007), the UIUC Sports Event dataset (Li and Fei-Fei 2007), and the MIT Scene dataset (Oliva and Torralba 2001). Specifically, the H-descriptor is first assessed in seven different color spaces, and then the H-fusion descriptor is compared with other popular descriptors, such as combinations of Scale Invariant Feature Transform (SIFT) (Lowe 1999, 2004) with other descriptors (Li and Fei-Fei 2007; Bo et al. 2011), the Pyramid Histograms of visual Words (PHOW) descriptor (Bosch et al. 2007a), LBP (Ojala et al. 1994) and Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b) based descriptors (Banerji et al. 2011), Spatial Envelope (Oliva and Torralba 2001), Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), and other approaches such as Object Bank (Li et al. 2010) and Hierarchical Matching Pursuit (Bo et al. 2011).

6.2.1 Datasets Used

This section briefly describes the three publicly available and fairly challenging image datasets. All of these datasets are widely used for evaluating the performance of object and scene image descriptors and classification methods.

The Caltech 256 Dataset

The Caltech 256 dataset (Griffin et al. 2007) holds 30,607 images divided into 256 object categories and a clutter class. This dataset has been described in detail in Section 4.4.1. Some sample images from this dataset are shown in Figure 4.6(a), which reveals that while some classes like bear and teddy-bear have similar foreground objects, elements like the American flag and people are present in the background of many categories and hence their inter-class variability is low.

The UIUC Sports Event Dataset

The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains eight sports event categories. This dataset has been described in detail in Section 4.4.1. Some sample images are displayed in Figure 4.6(b).

The MIT Scene Dataset

The MIT Scene dataset, also known as the OT Scenes dataset (Oliva and Torralba 2001) has 2,688 images divided into eight categories. A detailed description of this dataset is provided in Section 3.4.1. Figure 3.4(a) shows a few sample images from this dataset.

6.2.2 Comparative Assessment of the H-descriptor in Seven Different Color Spaces

Now the H-descriptor is assessed in seven different color spaces — the RGB, oRGB, HSV, YIQ, YCbCr, $I_1I_2I_3$, and DCS color spaces — for image classification performance using the three datasets. Note that some large scale images are resized so that the larger dimension is reduced to 400 pixels. To derive the H-descriptor from each image, the 3D-LBP descriptor is first computed to produce three new color images. Then the Haar wavelet transform of the 3D-LBP color images and the original color image is calculated. The HOG descriptors are further computed from the Haar wavelet transform of the component

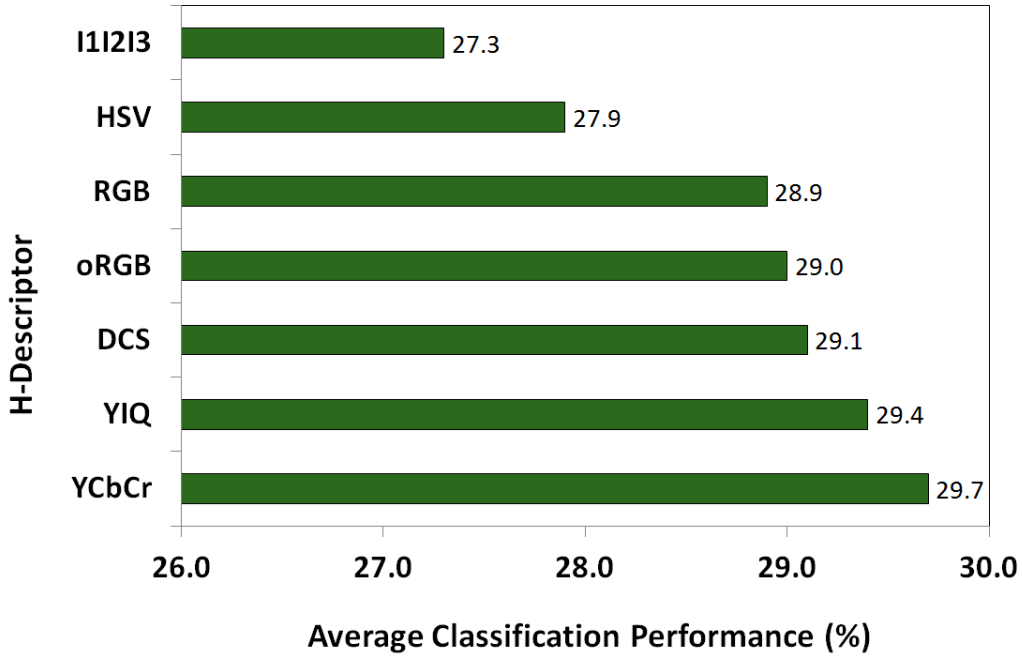


Figure 6.3 The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, RGB, oRGB, DCS, YIQ, and YCbCr color spaces using the EFM-NN classifier on the Caltech 256 dataset.

images of the color image and its 3D-LBP color images. Finally, the H-descriptor is derived by concatenating the HOG descriptors of the Haar wavelet transformed color images. Each image is transformed in the seven color spaces and the same operations are performed to construct the seven different color H-descriptors. Next, PCA is applied to reduce the dimensionality of the H-descriptors to derive the most expressive features, which are further processed by EFM to obtain the most discriminatory features for classification, and finally the nearest neighbor rule is used for image classification.

For the Caltech 256 dataset, a protocol defined in (Griffin et al. 2007) is used. On this dataset, experiments are conducted for the H-descriptors from seven different color spaces. For each class, 50 images are used for training and 25 images for testing. The data splits are the ones that are provided on the Caltech website (Griffin et al. 2007). Figure 6.3 shows the detailed performance of the H-descriptors using the EFM-NN classifier

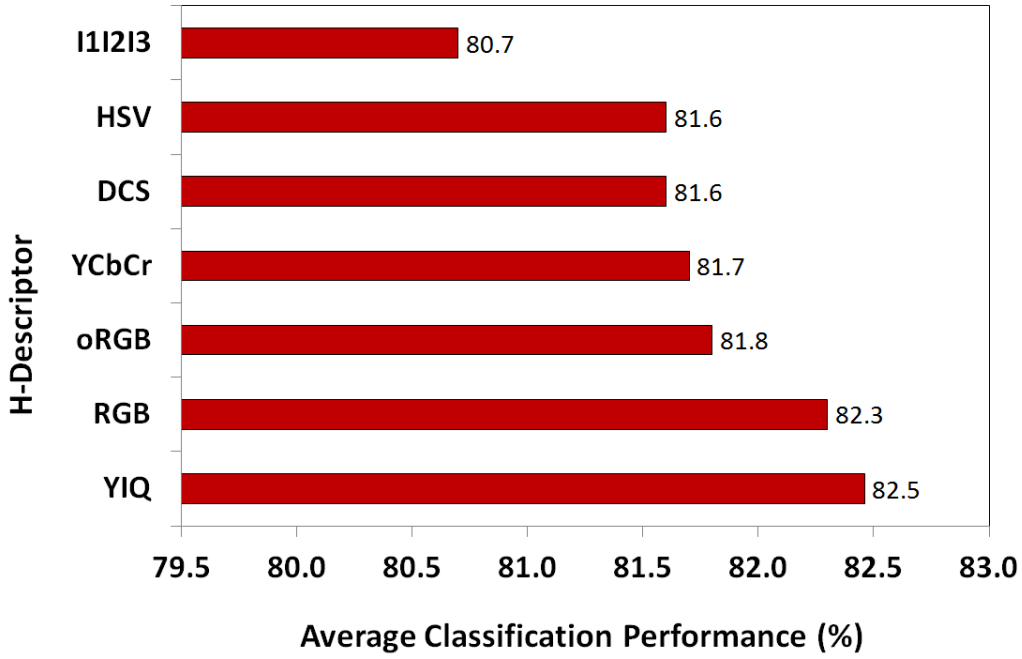


Figure 6.4 The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, DCS, YCbCr, oRGB, RGB, and YIQ color spaces using the EFM-NN classifier on the UIUC Sports Event dataset.

on the Caltech 256 dataset. The horizontal axis indicates the average classification performance, which is the percentage of correctly classified images averaged across the 256 classes and the five runs of the experiments, and the vertical axis shows the seven different H-descriptors in the seven color spaces. Among the different H-descriptors, the H-descriptor in the YCbCr color space achieves the best average classification performance of 29.7%, followed by the H-descriptors in the YIQ, DCS, oRGB, RGB, HSV and $I_1I_2I_3$ color spaces with the average classification performance of 29.4%, 29.1%, 29.0%, 28.9%, 27.9%, and 27.3%, respectively.

For the UIUC Sports Event dataset, a protocol defined in (Li and Fei-Fei 2007) is used, which specifies that for each class in this dataset, 70 images are used for training and 60 images for testing. To achieve more reliable performance, the experiments are repeated five times using random splits of the data, and no overlapping occurs between the training

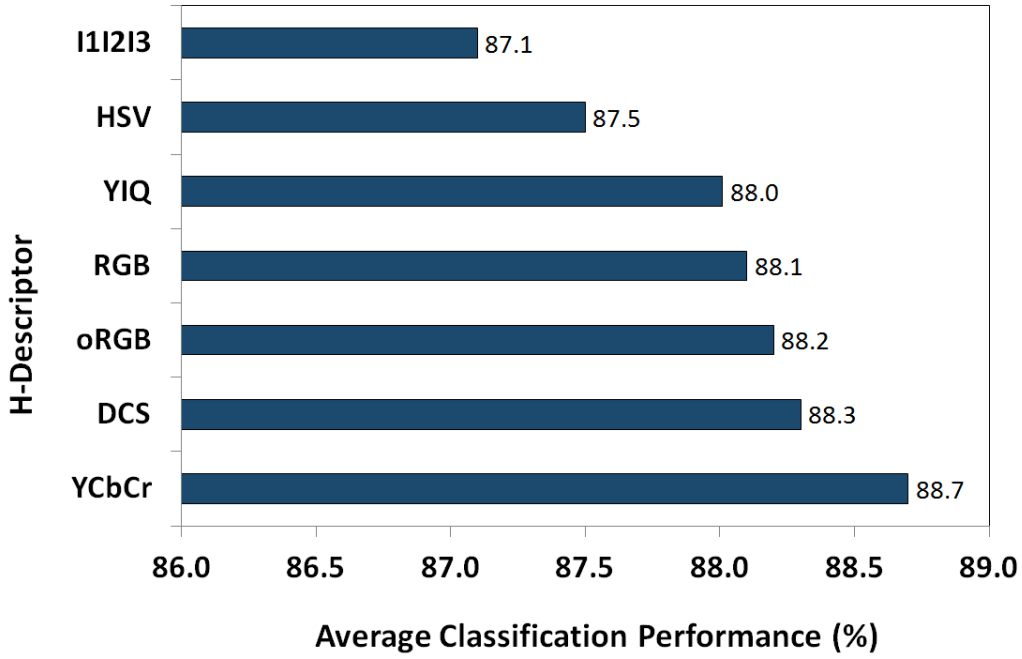


Figure 6.5 The average classification performance of the proposed H-descriptor in the $I_1I_2I_3$, HSV, YIQ, RGB, oRGB, DCS, and YCbCr color spaces using the EFM-NN classifier on the MIT Scene dataset.

and the testing sets of the same split. Figure 6.4 shows that the H-descriptor in the YIQ color space is the best descriptor with 82.5% average classification performance followed in order by the H-descriptors in the RGB, oRGB, YCbCr, DCS, HSV and $I_1I_2I_3$ color spaces with 82.3%, 81.8%, 81.7%, 81.6%, 81.6% and 80.7% success rates, respectively. Again the horizontal axis indicates the average classification performance and the vertical axis the H-descriptors in the seven color spaces.

For the MIT Scene dataset, 250 images are used from each class for training and the rest of the images for testing. All experiments are performed for five random splits of the data. Figure 6.5 reveals that the H-descriptor in the YCbCr color space performs the best with 88.7% average classification rate. The H-descriptors in the DCS, oRGB, RGB, YIQ, HSV and $I_1I_2I_3$ color spaces correctly classify on an average 88.3%, 88.2%, 88.1%, 88.0%, 87.5% and 87.1% of the images respectively. Again the horizontal axis shows the

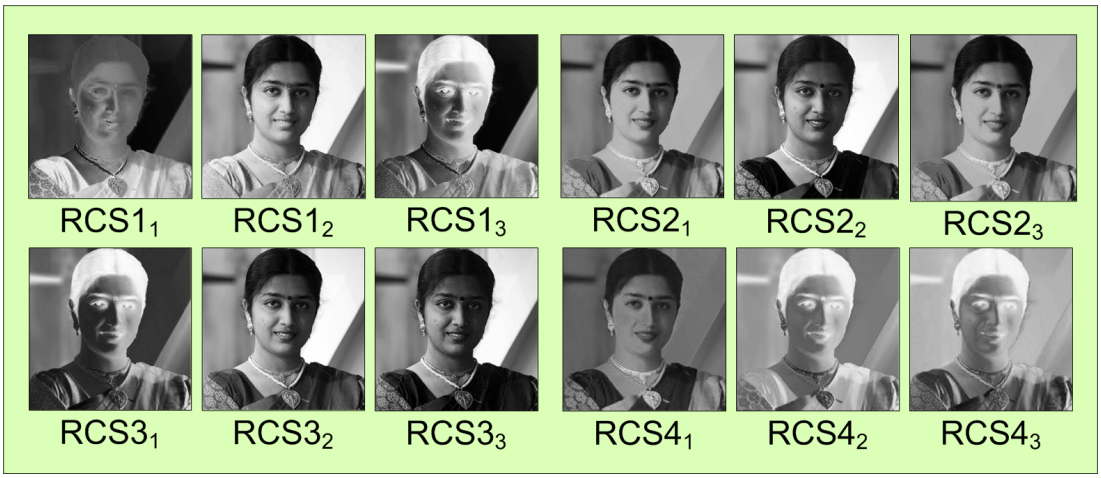


Figure 6.6 The color component images of the image from Figure 2.1 in the four random color spaces, namely RCS1, RCS2, RCS3 and RCS4 color spaces, respectively.

average classification performance and the vertical axis the H-descriptors.

6.2.3 Random Color Spaces and Performance of the H-descriptor in These Color Spaces

To further establish the robustness of the proposed H-descriptor for object and scene image classification, four random color spaces are generated and the classification performance is assessed using the descriptor in these color spaces. To generate a random color space, a 3×3 transformation matrix is created with randomly chosen elements.

$$\begin{bmatrix} R_1 \\ R_2 \\ R_3 \end{bmatrix} = \begin{bmatrix} W_{11} & W_{12} & W_{13} \\ W_{21} & W_{22} & W_{23} \\ W_{31} & W_{32} & W_{33} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{6.1}$$

where R_1, R_2 and R_3 are the three color components in the new random color space, and $W_{ij} \in (-1,1)$ are pseudorandom numbers. The three color components in the new color

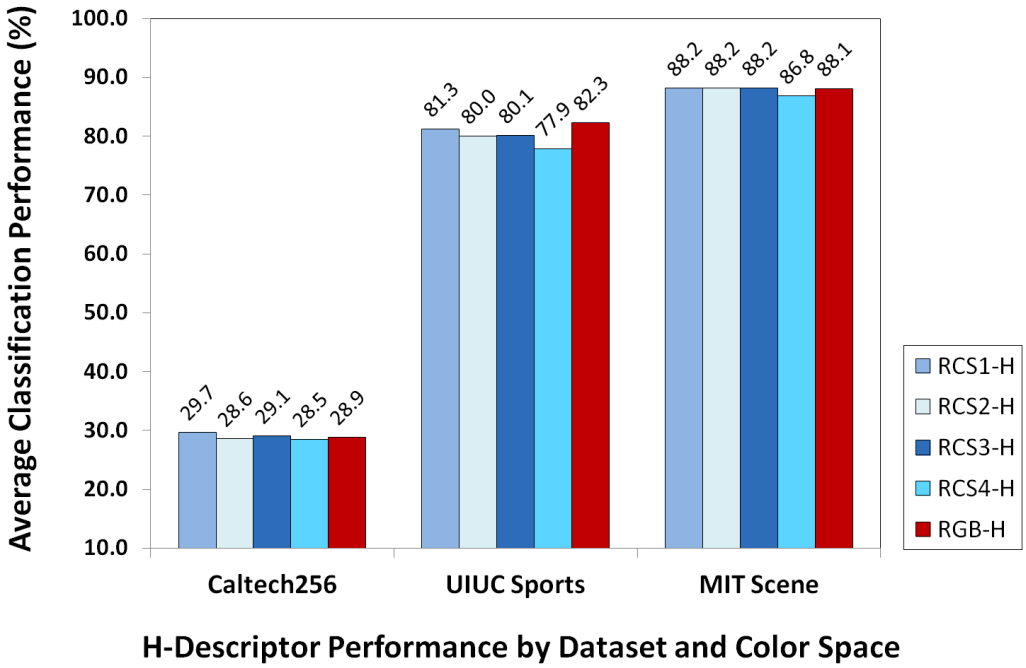


Figure 6.7 A comparison of the average classification performances of the H-descriptor in the RGB color space and the four random color spaces RCS1, RCS2, RCS3, and RCS4 on the three image datasets. Note that all the five descriptors apply the EFM-NN classifier.

space are thus given by

$$\begin{aligned} R_1 &= W_{11}R + W_{12}G + W_{13}B \\ R_2 &= W_{21}R + W_{22}G + W_{23}B \\ R_3 &= W_{31}R + W_{32}G + W_{33}B \end{aligned} \tag{6.2}$$

The classification performance of the proposed H-descriptor is next assessed in four random color spaces. In particular, four such random transformation matrices are generated and the resulting color spaces are named random color spaces 1, 2, 3 and 4 (RCS1, RCS2, RCS3 and RCS4). The original images from each of the three datasets mentioned above are transformed into each of these color spaces and subsequently the H-descriptor is generated, the same training and testing framework is used as for the other seven color spaces. Figure 6.6 shows the component images of the color image shown

before in Figure 2.1 in the four random color spaces used for these experiments. It should be noted that the images shown here are just four instances of what the components of a random color space could look like.

The results of the classification experiments are shown in Figure 6.7 with the performance in the RGB color space for reference. Here, the horizontal axis shows the H-descriptors in different color spaces, and the different datasets while the vertical axis shows the average classification performance. The performance of the H-descriptor in RCS1, RCS2, RCS3 and RCS4 remains, in all cases except one, within 2% of the performance of the H-descriptor in the RGB color space. First, these results show that the performance is random and unpredictable. In some cases it is more than the RGB H-descriptor performance and in other cases it is less. This indicates that simply transforming the color space does not increase the performance — the exact nature of the transformation is also important. Second, the results demonstrate that for the H-descriptors in RCS1, RCS2, RCS3 and RCS4 color spaces, the classification success rates stay reasonably close to the classification rate of the H-descriptor in the RGB color space. This indicates that the proposed descriptor is robust enough to yield stable performance under unpredictable changes in the color component values.

6.2.4 Comparative Assessment of the Grayscale H-descriptor, the Color H-descriptors and the H-fusion Descriptor

In this section, an attempt is made to investigate the importance of using color information for classification, and then to justify the fusion of H-descriptors in the seven different color spaces to form the H-fusion descriptor. Towards that end, a grayscale H-descriptor is generated and its classification performance is comparatively evaluated with the RGB H-descriptor and H-fusion descriptor.

The 3D-LBP operation, which is the first step of generating the H-descriptor, is only defined for a color image, i.e. an image with three component planes. This is because the

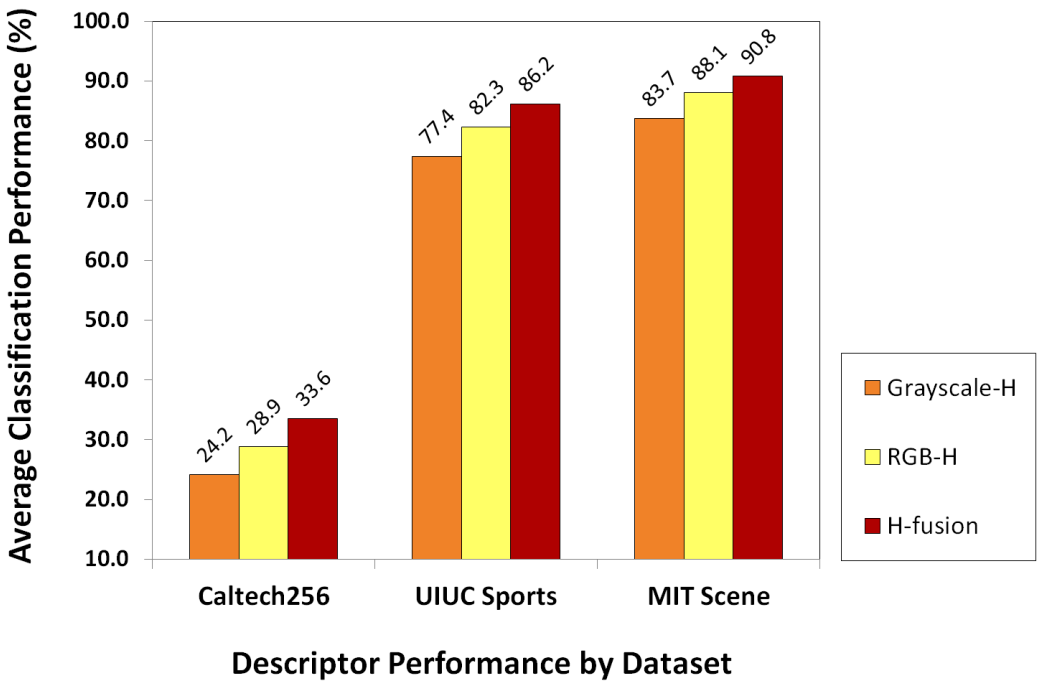


Figure 6.8 A comparison of the average classification performances of the H-descriptor in grayscale, in the RGB color space and the H-fusion descriptor on the three image datasets. Note that all the three descriptors apply the EFM-NN classifier.

3D-LBP captures the variations in pixel intensities across the color planes thus encoding image color information. To generate the H-descriptor for a grayscale image, it first needs to be converted to a three-plane image. In particular, for this experiment each color image with three planes is taken and converted to a grayscale image with just one plane by forming a weighted sum of the R, G, and B components as defined by Equation 2.9. Then that single plane is replicated twice to form a three-plane image again. The H-descriptor is subsequently generated from this image and classification performed using the EFM-NN classifier.

To create the H-fusion descriptor, the H-descriptor is computed from each image in each of the seven well-defined color spaces as described in Section 6.2.2. Then, after reducing the dimensionality of each of these seven feature vectors to $min(2000, rank - 1)$ PCA features, they are concatenated and form the H-fusion descriptor. Subsequently, the

dimensionality is further reduced using PCA and the most discriminatory features extracted using EFM. Here also, the final number of features before classification is one less than the number of categories.

Figure 6.8 compares the classification performance of the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor. Specifically, for the Caltech 256 dataset, the grayscale H-descriptor yields a success rate of 24.2%. Simply including RGB color information takes the correct classification rate up to 28.9% and the fusion of color spaces increases this by a further 4.7% to correctly classify 33.6% of the images. On the UIUC Sports Event dataset, the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor show classification rates of 77.4%, 82.3% and 86.2% respectively, thus demonstrating a significant advantage of using color. For the MIT Scene dataset, the classification rates obtained for the grayscale H-descriptor, the RGB H-descriptor and the H-fusion descriptor are 83.7%, 88.1%, and 90.8% respectively. Thus the H-fusion descriptor increases classification performance by over 7% from the grayscale H-descriptor, which is a quite high improvement for a dataset of this size and complexity. It should be noted that for the MIT Scene dataset, 250 images from each class are used for training in these experiments.

On comparing Figure 6.8 with Figure 6.3, Figure 6.4, Figure 6.5, and Figure 6.7, it is found that the classification performance of the grayscale H-descriptor is not only less than the RGB H-descriptor, but it is also less than the classification performance of the H-descriptor in any other color space as well. This is in accordance with the principle behind the 3D-LBP operation which is the first step of generating the H-descriptor. The 3D-LBP operation has been designed specifically to extract color information from the difference in pixel values in the three color component images, and since this difference is zero in a grayscale image, the H-descriptor does not perform as well for grayscale images as it does for color images. Also, the H-fusion descriptor performs better than the H-descriptor in any of the individual color spaces which justifies the fusion of H-descriptors from different color spaces.

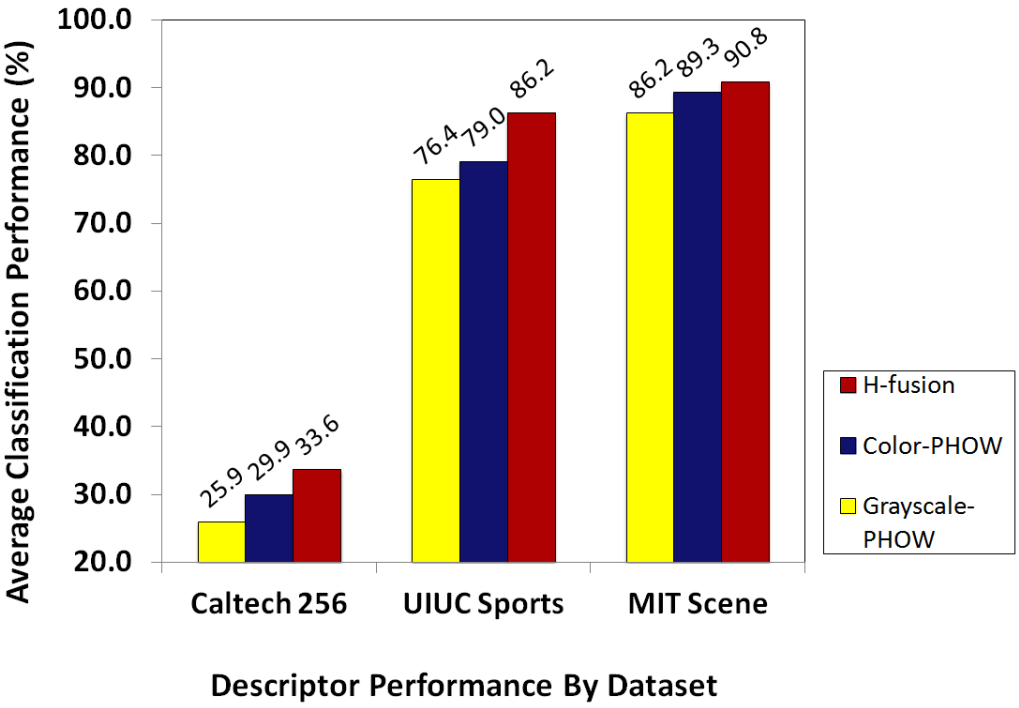


Figure 6.9 A comparison of the average classification performances of the color-PHOW descriptor, the grayscale-PHOW descriptor, and the proposed H-fusion descriptor on the three image datasets. Note that all the three descriptors apply the EFM-NN classifier.

6.2.5 Comparative Assessment of the H-fusion Descriptor and Some Popular State-of-the-art Image Descriptors

In this section the performance of the proposed H-fusion descriptor on the three datasets described in Section 6.2.1 is evaluated. First the proposed H-fusion descriptor is compared with the popular and robust SIFT-based Pyramid Histograms of visual Words (PHOW) descriptor (Bosch et al. 2007a). For fair comparison, both descriptors apply the EFM-NN classifier for image classification. Then the H-fusion descriptor is compared with some other popular state-of-the-art descriptors using the image classification performance reported in the published papers.

To make a comparative assessment of the H-fusion descriptor with a popular SIFT-based descriptor, the Pyramid Histograms of visual Words (PHOW) feature vector (Bosch et al. 2007a) is generated using the software package VLFeat (Vedaldi and Fulkerson 2010).

Here feature extraction is a three-step process. First SIFT features are extracted from images using a fast SIFT process. In this algorithm, SIFT descriptors are computed at points on a dense regular grid instead of the SIFT-generated interest points (Lazebnik et al. 2006; Bosch et al. 2007a). Next, the SIFT features are subjected to K-means clustering with K=1000 to form a visual vocabulary. Finally, the images are spatially tiled into 2×2 parts and the histograms of visual words are computed for the SIFT features from each part. These four histograms are concatenated to generate the final PHOW feature vector. For a color image, the same process is repeated for the three color component images and the feature vectors are concatenated. The grayscale PHOW and the color PHOW feature vectors are coupled with the EFM-NN classifier to compare the classification performance. Please note that the SIFT process applied here is an optimized C code that is 30 to 70 times faster than the conventional SIFT method (Vedaldi and Fulkerson 2010). In comparison, the H-descriptor is implemented using the MATLAB code that is not optimized in terms of computational efficiency. However, the vector generation time for the color PHOW is slightly longer than that for the color H-descriptor. For both PHOW and H-fusion descriptors, PCA is applied for dimensionality reduction and EFM-NN is applied for classification in order to make a fair comparison.

Figure 6.9 shows that the proposed H-fusion descriptor has an image classification performance better than both the grayscale and the color PHOW descriptors on the Caltech 256 dataset. Note that the horizontal axis of this graph lists the three descriptors and the three datasets while the vertical axis shows the average classification performance as a percentage. In particular, the H-fusion descriptor achieves the average classification rate of 33.6%, compared to the color-PHOW descriptor with the average classification rate of 29.9% and to the grayscale-PHOW descriptor with the average classification rate of 25.9%, respectively. Please note that the classification performance for the Caltech 256 dataset is quite low, because this dataset has a very high intra-class variability and in several cases the object occupies a small portion of the full image.

Table 6.1 Comparison of the Classification Performance (%) of the H-fusion Descriptor with other Popular Methods on the UIUC Sports Event Dataset

	#train = 560, #test = 480	
H-fusion	**Proposed Descriptor**	**86.2**
HMP	(Bo et al. 2011)	85.7
SIFT+SC	(Bo et al. 2011)	82.7
OB	(Li et al. 2010)	76.3
SIFT+GGM	(Li and Fei-Fei 2007)	73.4

Figure 6.9 also displays the image classification performance on the UIUC Sports Event dataset. Specifically, the H-fusion descriptor correctly classifies 86.2% of the images and performs better than both the grayscale and the color PHOW descriptors, which achieve the average classification performance of 76.4% and 79.0%, respectively. Using this UIUC Sports Event dataset, the H-fusion descriptor is further compared, with some popular state-of-the-art descriptors and methods, such as the Hierarchical Matching Pursuit (Bo et al. 2011), Object Bank approach (Li et al. 2010) and variations of the popular Scale Invariant Feature Transform (SIFT) (Lowe 2004) descriptor (Bo et al. 2011; Li and Fei-Fei 2007). Note that the performance reported here for the competing methods are from the published papers. Table 6.1 shows that the H-fusion descriptor achieves the best classification performance of 86.2% compared to HMP (Bo et al. 2011) with classification performance of 85.7%, to SIFT+SC(Bo et al. 2011) with classification performance of 82.7% , to Object Bank (Li et al. 2010) with classification performance of 76.3% and to the SIFT+GGM (Li and Fei-Fei 2007) method with classification performance of 73.4%.

On the MIT Scene dataset, two sets of experiments are performed with the H-fusion descriptor. First 250 images from each class are used for training and the rest of the images for testing. In this set of experiments, the proposed H-fusion descriptor yields an average success rate of 90.8% and exceeds the performance achieved by the PHOW descriptors. Figure 6.9 shows the image classification performance on this dataset as well. Specifically, the H-fusion descriptor correctly classifies 90.8% of the images and performs better

Table 6.2 Comparison of the Classification Performance (%) of the H-fusion Descriptor with other Popular Methods on the MIT Scene Dataset

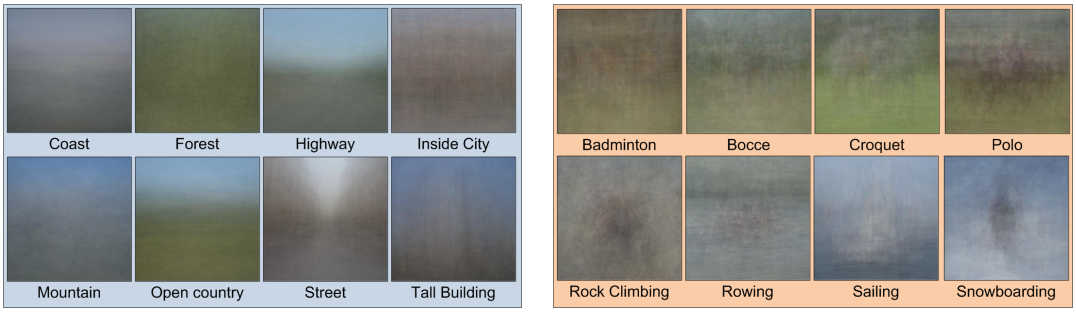
	#train = 2000, #test = 688	
H-fusion	**Proposed Descriptor**	**90.8**
CGLF+PHOG	(Banerji et al. 2011)	89.5
CGLF	(Banerji et al. 2011)	86.6
PHOG	(Banerji et al. 2011)	79.1
	#train = 800, #test = 1888	
H-fusion	**Proposed Descriptor**	**87.7**
C4CC	(Bosch et al. 2006)	86.7
CGLF+PHOG	(Banerji et al. 2011)	84.3
SE	(Oliva and Torralba 2001)	83.7
CGLF	(Banerji et al. 2011)	80.0

than both the grayscale and the color PHOW descriptors, which achieve the average classification performance of 86.2% and 89.3%, respectively. In the next set of experiments 100 images are used from each class for training and the remaining images for testing. The proposed descriptor is further compared with some widely used state-of-the-art descriptors and classification approaches such as the Spatial Envelope (Oliva and Torralba 2001), Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006), Color Grayscale LBP Fusion (CGLF) (Banerji et al. 2011) and Pyramid Histograms of Oriented Gradients (PHOG) (Bosch et al. 2007b; Banerji et al. 2011). Here also, the results achieved by other researchers are reported directly from their published work. Table 6.2 shows that with 250 training images, the proposed H-fusion descriptor achieves the best classification performance of 90.8% as compared to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 89.5%, to CGLF (Banerji et al. 2011) with a classification performance of 86.6% and to PHOG (Bosch et al. 2007b; Banerji et al. 2011) with a classification performance of 79.1%. With 100 training images per class, the H-fusion descriptor again yields the best classification performance of 87.7%, as compared to Color SIFT four Concentric Circles (C4CC) (Bosch et al. 2006) with a classification performance of 86.7%, to CGLF+PHOG (Banerji et al. 2011) with a classification performance of 84.3%, to Spatial

Envelope with a classification performance of 83.7%, and to CGLF (Banerji et al. 2011) with a classification performance of 80.0%.

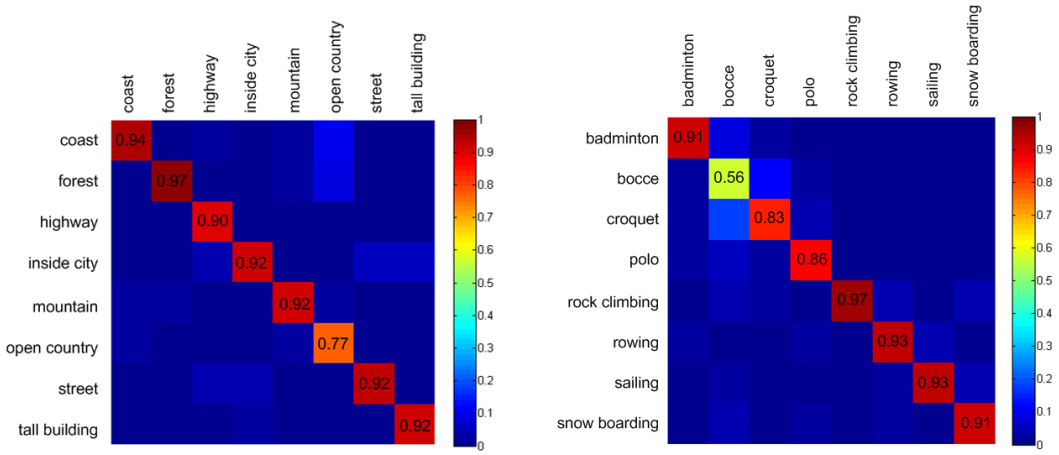
6.3 Discussion

The descriptors proposed in this chapter have been thoroughly tested for classification performance on several datasets. They are very different in their image properties. While the KTH-TIPS and KTH-TIPS2-b datasets used in Chapter 3 have been created by photographing textures in a lab under controlled lighting conditions, the three datasets used in this chapter, namely the Caltech 256 dataset, the MIT Scene dataset, and the UIUC Sports Event dataset, are composed of images collected from the Internet. Among these, the MIT Scene dataset images are all color photographs that have been standardized to a 256×256 pixel size. The UIUC Sports Event dataset images, on the other hand, are highly variable in size and the mean image length is over 1000 pixels which necessitates resizing before feature extraction. This dataset also contains a few grayscale images. The Caltech 256 dataset is the most complex dataset used, with both color and grayscale images, and even non-photographic images. This section tries to take a look at some of these datasets and further analyze the experimental results to better understand the performance, beyond the numbers.



(a) (b)

Figure 6.10 The category mean images from the (a) MIT Scene dataset and (b) UIUC Sports Event dataset.



(a) (b)

Figure 6.11 The confusion matrices for classification by the H-fusion descriptor and the EFM-NN classifier for the (a) MIT Scene dataset and (b) UIUC Sports Event dataset.

The mean image of a category in an image dataset is one measure of intra-class image variability. Figure 6.10 shows the mean images from all categories of the MIT Scene dataset and the UIUC Sports Event dataset in attempt to demonstrate the relative intra-class variability of the various categories. In particular, Figure 6.10(a) shows that some categories from the MIT Scene dataset like coast, open country and street have a discernible structure in their mean image while means of categories like forest and highway have distinctive colors. The inside city class has the most blurred mean. The fact that the proposed H-fusion descriptor classifies almost all the classes equally well demonstrates the robustness of this image descriptor. From Figure 6.10(b), it can be seen that the mean of the categories in the UIUC Sports Event dataset are also largely blurred although some amount of structure is visible in snowboarding, sailing, polo and croquet classes. The presence of similar backgrounds in different categories makes this dataset so challenging.

Next, an analysis is done of the category-wise classification rates on the MIT Scene and the UIUC Sports Event Dataset. The success rates given in this section are based on the H-fusion descriptor. Figure 6.11(a) shows a confusion matrix for classification using the H-fusion descriptor on the MIT Scene dataset with the categories in alphabetical order.

Figure 6.11(b) shows a similar confusion matrix for the UIUC Sports Event dataset. In each confusion matrix, the rows show assigned classes while the columns show actual classes. For instance, a high value at row 1, column 6 signifies that a lot of images from class 6 (open country) get assigned the class label 1 (coast). In the experiments for the MIT Scene dataset, 250 images from each class were used for training and the remaining for testing. For the UIUC Sports Event dataset, 70 images from each class were used for training and 60 for testing.

It can be seen from Figure 6.11(a) that the best classified categories are forest and coast with success rates of 97% and 94% respectively. This is in accordance with the fact that the means of these categories are easy to identify. However, although the mean of the category open country seems to have an identifiable structure, it is the most difficult category to classify. As the confusion matrix shows, some of the open country scenes are classified as coast, some as forest and some as mountain scenes. This is because the class open country has been defined with a lot of overlap with other classes. Parts (a), (b) and (c) of Figure 6.12 show some of the particularly confusing images from the open country category that get misclassified as coast, forest and mountain respectively. The other three

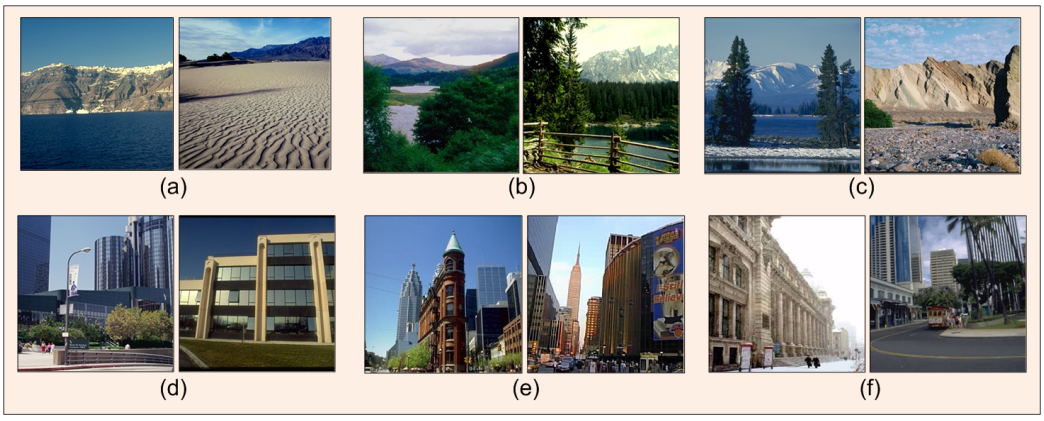


Figure 6.12 Some ambiguous images from the MIT Scene dataset. Parts (a), (b) and (c) show some images from the open country category that get misclassified as coast, forest and mountain respectively. Parts (d), (e) and (f) show ambiguous images from the inside city, tall building and street categories respectively that contain similar features.

Figure 6.13 Some images from the UIUC Sports Event dataset showing the high intra-class and low inter-class variance. Part (a) shows some images from the bocce class which has the lowest classification rate and part (b) shows images from the croquet, rock climbing, badminton, sailing and polo classes. These are the classes where most of the wrongly classified images from the bocce class are classified.

categories that are confused with each other are inside city, street and tall buildings. Parts (d), (e) and (f) of Figure 6.12 show two images each from the inside city, tall building and street categories respectively that contain similar elements and hence cause misclassification. These results are similar to those reported by (Oliva and Torralba 2001) which would indicate that the confusion is due to an inherent ambiguity in the manual annotation of these particular dataset categories themselves.

Figure 6.11(b) shows the category-wise classification performance on the UIUC Sports Event dataset. From this confusion matrix, it can be seen that the worst classification performance here occurs with the bocce class which has the maximum intra-class variation. Figure 6.13(a) shows some images from the bocce class and Figure 6.13(b) shows similar images from the classes where the wrongly classified bocce images have been mostly placed. Note the diversity of the background in the bocce images.

The Caltech 256 dataset is much more complex and varied in its composition of categories and hence it is difficult to explain the classification performance on this dataset using one-to-one category misclassifications. The top row of Figure 6.14 shows the cate-

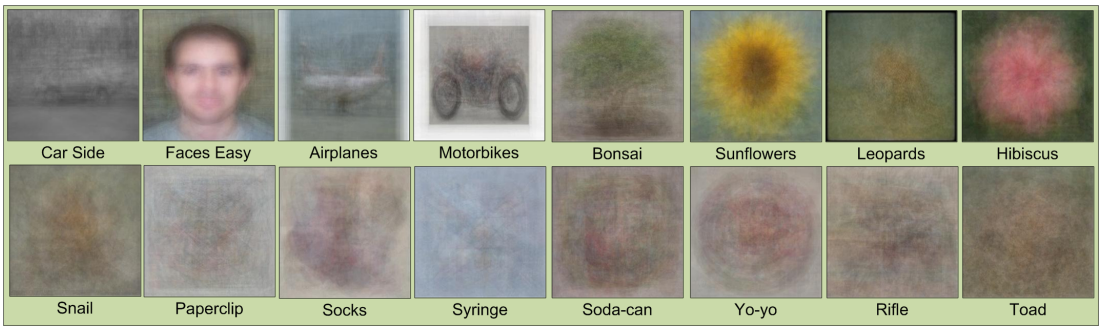


Figure 6.14 The category mean images from the eight most successful categories (upper row) and the eight least successful categories (lower row) from the Caltech 256 dataset.

gory means of the eight classes from this dataset which show the best classification performance. The bottom row of the same figure shows the means of the eight classes that are most difficult to classify. Almost all the images in the bottom row are uniformly blurred while for those in the top row, the categories are distinguishable from their means.

One significant characteristic of the Caltech 256 dataset is that many of the classes in this dataset are made based on semantic concepts rather than image characteristics. For

Figure 6.15 Some images from the drinking-straw category of the Caltech 256 image dataset showing the intra-class variability. Several classes in this dataset are based on semantic concepts rather than image characteristics.

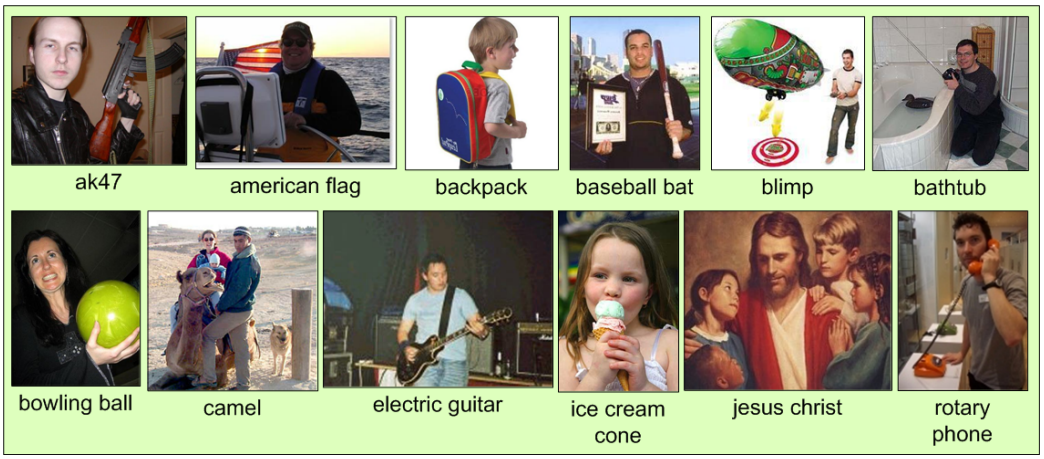


Figure 6.16 Some images from the Caltech 256 image dataset. None of the images are from the people class although all contain human figures. The categories each image belongs to is indicated below the image.

instance, Figure 6.15 shows a few images from the drinking straw category of this dataset. The images are vastly different from each other and in some cases, contain other elements that are occupy a more significant area of the image. Such classes show poor classification performance and bring down the overall classification average. Apart from high intra-class variation, low inter-class variation is also another problem with this dataset. Figure 6.16 shows some images from different classes that contain human figures. The point to be noted here is that although the human figures occupy a significant part of all of these images, none of them belong to the people class. In general, similar situations can be found in most classes where images of one class contain objects of another class. One possible course of action for future works could be a fuzzy class membership for each image where typically an image is assigned multiple class labels in order of probability. That way, a man holding a gun would be classified both as a man and a gun which would be a more logical way to classify the images in this dataset.

6.4 Summary

This chapter has presented new image descriptors based on color, texture, shape, and wavelets for object and scene image classification. In particular, it has first presented a novel H-descriptor, which integrates the 3D-LBP and the HOG of its Haar wavelet transform, to encode color, texture, shape, and local information. It has also comparatively assessed the H-descriptor in seven different color spaces — the RGB, the HSV, the YCbCr, the oRGB, the $I_1I_2I_3$, the YIQ, and the DCS color spaces — for image classification performance. Finally a new H-fusion descriptor has been presented by fusing the PCA features of the H-descriptors in the seven color spaces. Experimental results using three datasets show that the proposed new H-fusion descriptor achieves better image classification performance than other popular descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank, the Hierarchical Matching Pursuit, as well as LBP. Finally, a detailed discussion of the experimental results has been included to show the category wise classification performance of the H-Fusion descriptor on the different datasets, and to highlight some characteristics of these three image datasets.

CHAPTER 7

BOWL: A NEW APPROACH TO USING LOCAL BINARY PATTERNS

In Chapter 3, the LBP descriptor was introduced and a novel mLBP descriptor was proposed to encode the texture of an image. The experiments with the KTH-TIPS, KTH-TIPS2-b and the MIT Scene datasets demonstrated the proposed descriptor to be good for scene and texture image classification. The mLBP descriptor uses three LBP neighborhoods to create a histogram of pixel patterns present throughout the image.

In recent times, however, a lot of researchers have obtained very promising results with part-based methods (Fei-Fei and Perona 2005; Csurka et al. 2004). Here the image is considered as a collection of sub-images or patches and the feature describes each part and not the whole image. Finally, similar parts are clustered together and a histogram of the parts, rather than the raw features, is used to represent the image. This approach is known as a "bag-of-words model", with each part representing a "visual word" that describes a part of the whole scene (Yang et al. 2007; Jiang et al. 2007).

This chapter explores a new bag-of-words based image descriptor that makes use of the multi-mask LBP concepts introduced in Chapter 3, but significantly improves the classification rate. Bag-of-words models work well with spatial pyramid representations and Support Vector Machine (SVM) classifiers (Zhang et al. 2010) and so both concepts have been used for the experiments. Experiments with three publicly available datasets, namely the MIT Scene dataset, the Fifteen Scenes dataset and the UIUC Sports Event dataset, show that the proposed Bag-of-words LBP (BoWL) descriptor not only improves classification performance over LBP and mLBP, but can also yield better results than other popular descriptors.

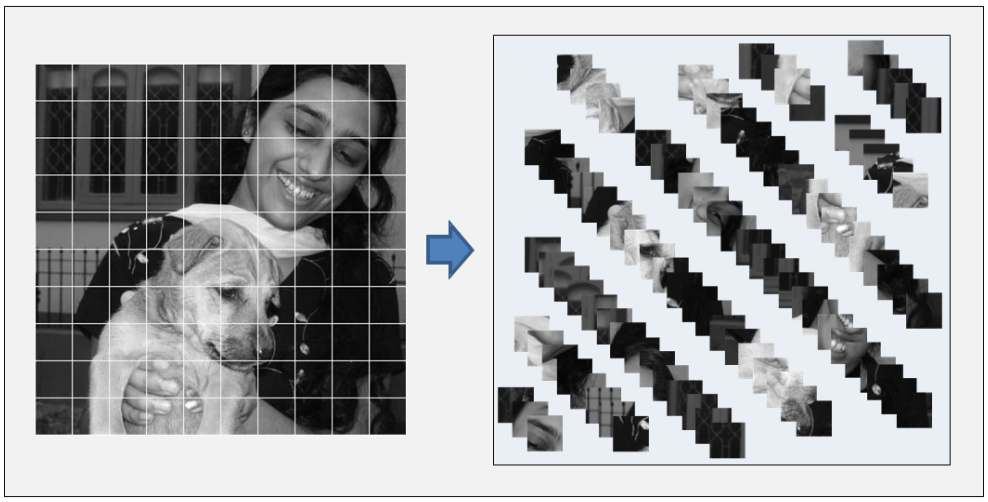


Figure 7.1 For the bag-of-words representation, a grayscale image is broken down into small image patches using a regular grid. This is called dense sampling. Overlapping patches are used for more accuracy.

7.1 An Innovative Bag of Words Local Binary Patterns Descriptor for Image Classification

This section explains in detail the various steps involved in computing the proposed BoWL descriptor from a grayscale image. These steps involve breaking down the image into small sub images, then extracting the features from these parts, and finally quantizing and forming the BoWL histogram for classification.

7.1.1 Formation of a Bag of Features from an Image

The first step in generating the new BoWL descriptor is the selection of small image patches. This process is known as sampling. Some image descriptors like SIFT (Lowe 2004) use multiscale keypoint detectors such as Laplacian of Gaussian or Harris-affine to select regions of interest within the image. While this sampling method is suitable for object recognition, it has been shown that dense or even random sampling often outperforms the keypoint-based sampling methods (Nowak et al. 2006). This is particularly true of scene images if the image has large uniform regions such as the sky, since no interest

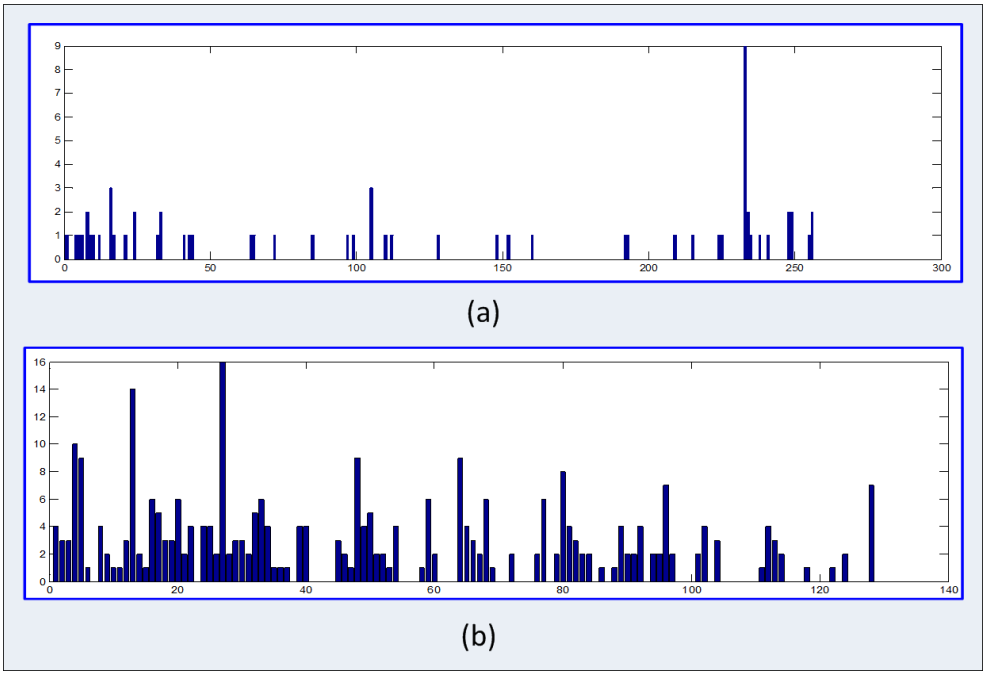


Figure 7.2 (a) Shows the traditional LBP histogram of a 10×10 pixel image patch. The vector is sparse and features are mostly similar. (b) shows the 128-component modified multi-neighborhood LBP vector of the same image patch obtained using the neighborhoods shown in Figure 7.3.

points are selected from those regions. The method proposed here uses dense feature extraction, which means the image is divided into a large number of equal sized blocks using a uniform grid and each block is used as a separate region for feature extraction. To increase classification performance, overlapping image patches are used. This process is explained in Figure 7.1. The image shown on the left is divided into uniform image patches by the regular grid displayed overlaid on the image, to form the image patches shown on the right. Such patches are created from all training images before the feature extraction is done.

7.1.2 A DCT-smoothed multi-mask LBP for Small Image Blocks

The original LBP image descriptor (Ojala et al. 2002) works by thresholding each pixel value based on the pixel values in its immediate neighborhood. Different researchers have experimented with styles of selecting the neighborhood, leading to different forms of the

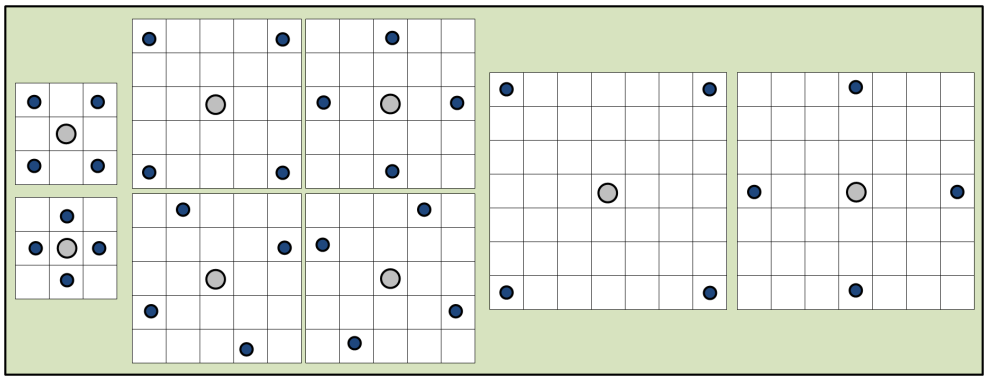


Figure 7.3 The eight neighborhoods for computing the modified LBP descriptors for small image patches. Please note that these neighborhoods are at different distances from the center pixel.

LBP descriptor. The neighborhood may even be selected using nearby features instead of geometric proximity to the pixel (Gu and Liu 2013). Figure 3.2 shows the three 8-pixel neighborhoods used for generating the mLBP descriptor introduced in Chapter 3. The LBP process assigns one out of 2^8 possible intensity values to each pixel. Thus the histogram produced in the subsequent step has 256 bins. However, if this technique is applied to a small image patch with ~256 pixels the histogram becomes sparse, with many bins having identical values. Figure 7.2(a) shows such a histogram for an image patch 10×10 pixels in size. To solve this problem, eight smaller neighborhoods of four pixels each are used. These eight neighborhoods produce a more dense 16-bin histogram, and eight such histograms from eight different neighborhoods are concatenated to generate the 128-dimensional feature vector describing each image patch. Figure 7.3 shows the 8 four-neighborhood LBP masks for generating the feature vector from each image patch. Please note that these neighborhoods are at different distances from the target pixel. Figure 7.2(b) shows the feature vector obtained by the modified LBP operation on the same image patch.

The Discrete Cosine Transform (DCT) can be used to transform an image from the spatial domain to the frequency domain, where an image is decomposed into a combination of various uncorrelated frequency components. Specifically, the DCT of an image with the spatial resolution of $M \times N$, $f(x,y)$, where $x = 0, 1, \cdots, M-1$ and $y = 0, 1, \cdots, N-$

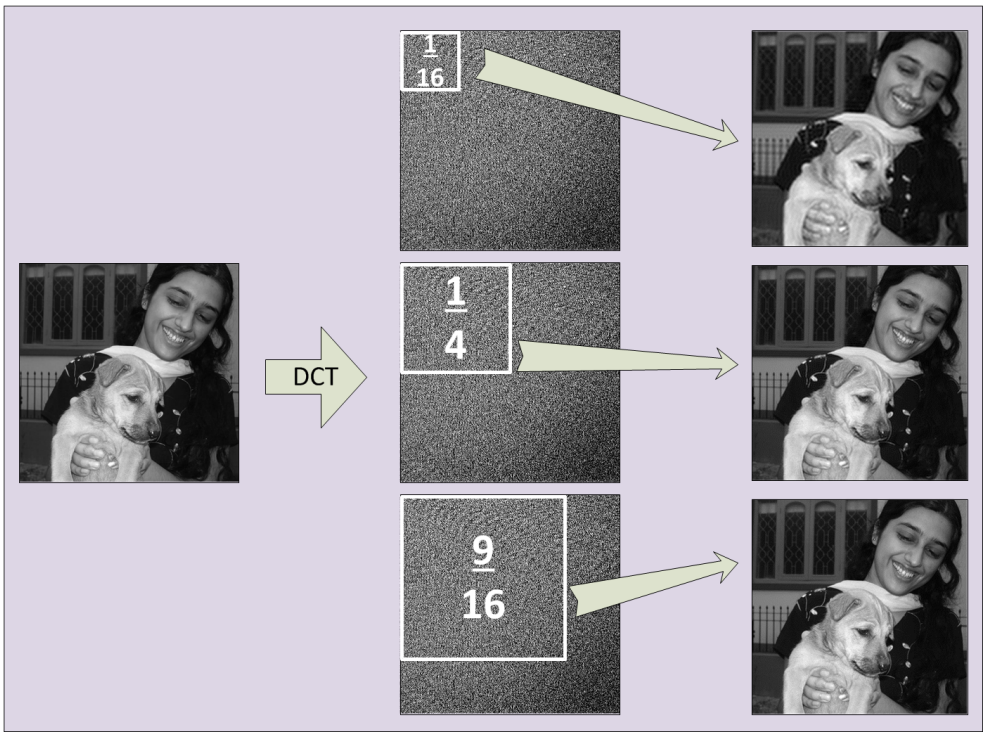


Figure 7.4 DCT can be used for smoothing out the image. The original image is transformed to the frequency domain and the lowest 1/16, 1/4 and 9/16 parts are used for regenerating the image, respectively, resulting in three output images with various degrees of smoothing.

1, transforms the image from the spatial domain to the frequency domain (Gonzalez and Woods 2008):

$$F(u,v) = \alpha(u)\alpha(v) \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y) cos \left[\frac{(2x+1)u\pi}{2M} \right] cos \left[\frac{(2y+1)v\pi}{2N} \right] \qquad (7.1)$$

where $\alpha(u) = \sqrt{1/M}$ for $u = 0$, $\alpha(u) = \sqrt{2/M}$ for $u = 1, 2, \cdots, M-1$, and $\alpha(v) = \sqrt{1/N}$ for $v = 0$, $\alpha(v) = \sqrt{2/N}$ for $v = 1, 2, \cdots, N-1$. DCT is thus able to extract the features in the frequency domain to encode different image details that are not directly accessible in the spatial domain. Due to these specific properties, DCT has been successfully applied to face recognition (Liu and Liu 2008; Chen et al. 2006; Hafed and Levine 2001). In the proposed method, DCT is used to eliminate higher frequencies from an image, resulting in a form of smoothing. Specifically, the original image is transformed to the frequency domain and

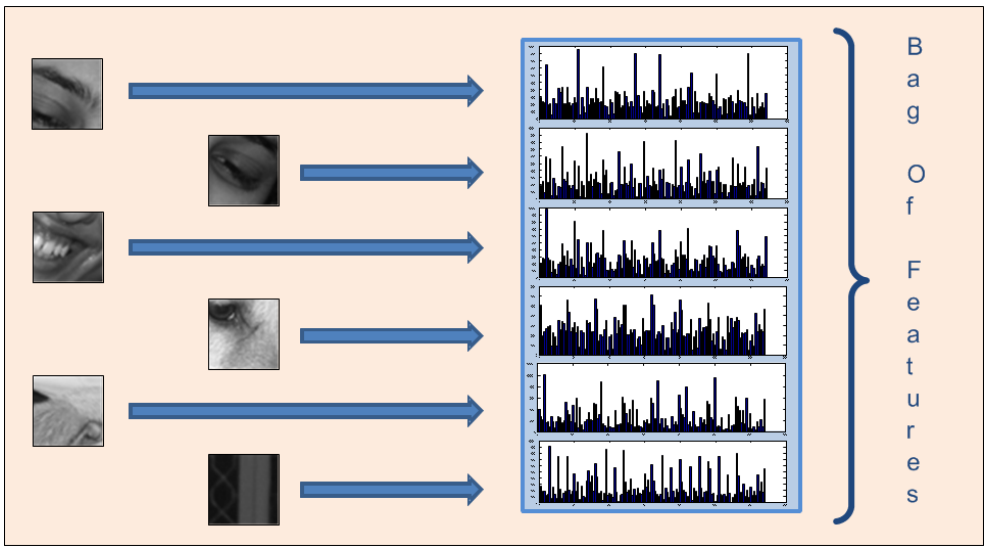


Figure 7.5 The features are computed from a large number of image patches from all training images and form a bag of features from which a visual vocabulary can be created.

the lowest 6.25%, 25% and 56.25% frequencies are used, respectively, for conversion back to the spatial domain to form three new images. This process is explained in Figure 7.4. The original image and the three images thus formed undergo the same process of dense sampling and eight-mask LBP. This means, effectively, the number of features extracted from an image is increased fourfold. All these features together form a bag of features, as shown in Figure 7.5, that needs to be clustered into distinct visual words to form a visual vocabulary.

7.1.3 Quantization, Pyramid Representation and Classification

As demonstrated in Figure 7.6, the bag of features extracted from the training images is next quantized into a visual vocabulary with discrete visual words. The popular k-means clustering method is used for this step. There is no consensus as to the proper size of a visual-word vocabulary. The vocabulary size used by other researchers varies from a few hundreds (Lazebnik et al. 2006; Zhang et al. 2007b), to several thousands and tens of thousands (Sivic and Zisserman 2003; Zhao et al. 2006). Their results are not directly compara-

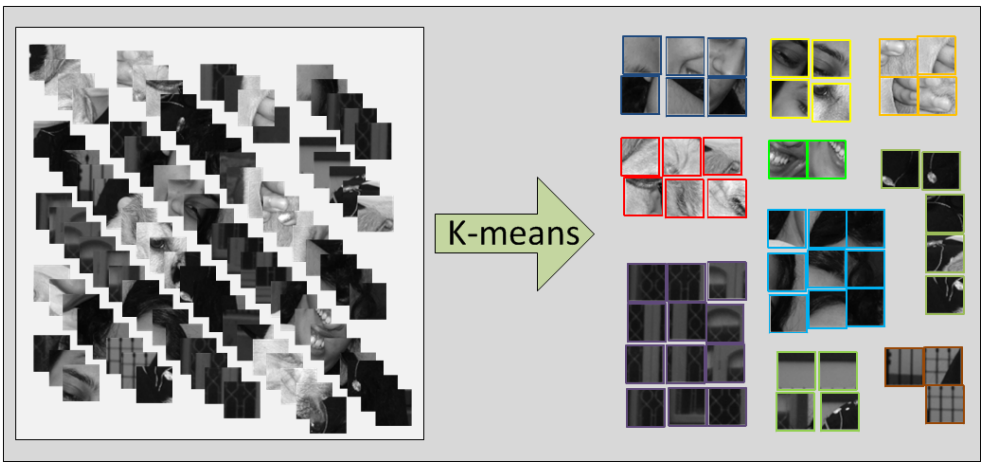


Figure 7.6 The features representing the small image patches are quantized into a number of visual words using a popular clustering method such as k-means to form a visual vocabulary.

ble due to different classification frameworks. To determine the right range of vocabulary size appropriate for the BoWL features, experiments were performed with vocabularies of sizes varying from 100 to 5,000. For the experiments presented in this dissertation, a 1000-word vocabulary was found to be optimum. After the formation of the visual vocabulary, each image patch from each training and test image is mapped to one specific word in the vocabulary. An image, therefore, can be represented by a histogram of visual words. This is explained in Figure 7.7(a).

Using the image pyramid representation of (Lazebnik et al. 2006), a descriptor is able to represent local image features and their spatial layout. In this method, an image is tiled into successively smaller blocks at each level and descriptors are computed for each block. The features from each level are weighted separately and all the features are finally concatenated to form a pyramid histogram. This technique is explained in Figure 7.7(b). It should be noted that the histograms shown in Figure 7.7 are for illustration purposes only. For this work, only the second level of this pyramid has been used to keep the computational complexity low. This creates a 4000 dimensional BoWL feature vector for each image.

After all training and test images have been processed and the feature vectors have

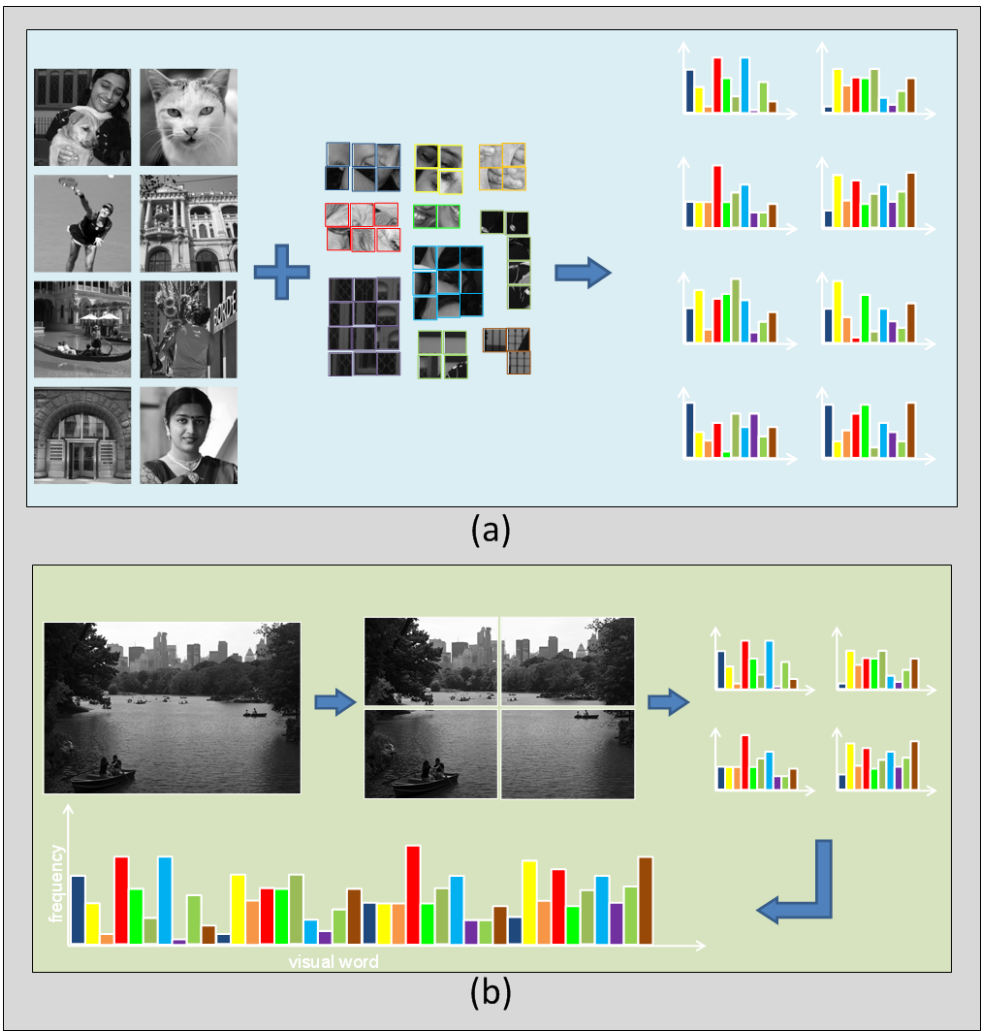


Figure 7.7 (a) All images are converted to histograms of visual words using the visual vocabulary created from the training images. (b) For the spatial pyramid representation, a full image is broken down into multiple spatial tiles. Then histograms of visual words are computed from each tile and concatenated.

been generated, an SVM classifier is used for classification. It is a known fact in texture and other image classification that for comparing histograms, using χ^2 or Hellinger distance measures usually yields better results than Euclidean distance (Arandjelović and Zisserman 2012) . The use of the Hellinger kernel has been shown to benefit SIFT (Arandjelović and Zisserman 2012). Since the proposed BoWL descriptor is also a histogram, intuitively it seems that it should yield better classification results with the Hellinger kernel and it is empirically seen that using the Hellinger kernel does indeed improve the classification

results greatly.

If x and y are n-vectors with unit Euclidean norm ($\|x\|_2 = 1$), then the Euclidean distance $d_E(x,y)$ between them is related to their similarity (kernel) $S_E(x,y)$ as

$$d_E(x,y)^2 = \|xy\|_2^2 = \|x\|_2^2 + \|y\|_2^2 - 2x^t y = 2 - 2S_E(x,y) \tag{7.2}$$

where $S_E(x,y) = x^t y$, and the last step follow from $\|x\|_2^2 = \|y\|_2^2 = 1$. The Euclidean similarity/kernel here needs to be replaced by the Hellinger kernel.

The Hellinger kernel, which is also known as the Bhattacharyya's coefficient, is defined for two L1 normalized histograms, x and y (i.e. $\sum_{i=1}^{n} x_i = 1$ and $x_i \geq 0$) as:

$$H(x,y) = \sum_{i=1}^{n} \sqrt{x_i y_i} \tag{7.3}$$

Arandjelović et al. suggest a simple algebraic manipulation to compare SIFT vectors by a Hellinger kernel (Arandjelović and Zisserman 2012). Since BoWL vectors are also based on histograms of words, the same technique can be applied to the BoWL vectors as well. This can be done in two steps: (i) L1 normalize the BoWL vector (originally it has unit L2 norm); (ii) square root each element. It then follows that $S_E(\sqrt{x}, \sqrt{y}) = \sqrt{x}^t \sqrt{y} = H(x,y)$, and the resulting vectors are L2 normalized since $S_E(\sqrt{x}, \sqrt{y}) = \sum_{i=1}^{n} = 1$ (Arandjelović and Zisserman 2012).

The key point is that comparing the square roots of the BoWL descriptors using Euclidean distance is equivalent to using the Hellinger kernel to compare the original BoWL vectors:

$$d_E(\sqrt{x}, \sqrt{y})^2 = 2 - 2H(x,y) \tag{7.4}$$

For the classification process, an SVM is trained independently for each class (one-vs-all classification). This is repeated for each category separately and the precision rates from all the iterations gives the average precision which is the mean classification accuracy. A similar configuration has been successfully used by other researchers like (Sanchez et al.

2012) in recent works. The SVM implementation used here is the one that is distributed with the VlFeat package (Vedaldi and Fulkerson 2010).

7.2 Experiments

This section first briefly introduces the three datasets used for testing the new BoWL image descriptor and then does a comparative assessment of the classification performance of the LBP, the mLBP and the BoWL descriptors. Finally the classification performance of the BoWL descriptor is compared with some other popular image descriptors used by other researchers.

7.2.1 Datasets Used

This section contains a brief overview of the three publicly available and widely used image datasets used for assessing the classification performance of the proposed descriptor.

The UIUC Sports Event Dataset

The UIUC Sports Event dataset (Li and Fei-Fei 2007) contains eight sports event categories. This dataset has been described in detail in Section 4.4.1.

The MIT Scene Dataset

The MIT Scene dataset, also known as the OT Scenes dataset (Oliva and Torralba 2001) has 2,688 images divided into eight categories. A detailed description of this dataset is provided in Section 3.4.1.

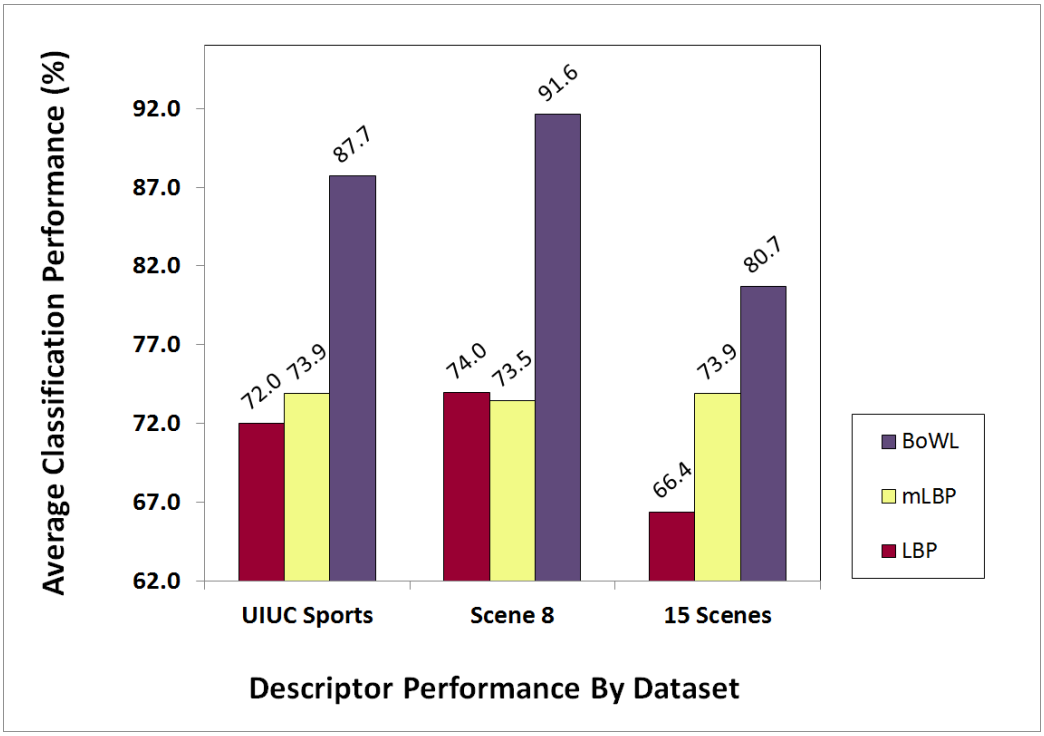


Figure 7.8 The mean average classification performance of the LBP, the mLBP and the proposed BoWL descriptors using an SVM classifier with a Hellinger kernel on the three datasets.

The Fifteen Scene Categories Dataset

The Fifteen Scene Categories dataset (Lazebnik et al. 2006) is composed of 15 scene categories. A detailed description of this dataset is provided in Section 4.4.1.

7.2.2 Comparative Assessment of the LBP, mLBP and BoWL Descriptors on the Different Datasets

In this section, a comparative assessment of the LBP, the mLBP and the proposed BoWL descriptor is made in grayscale, using the three datasets described earlier to evaluate classification performance. To compute the BoWL, first each training image is converted to grayscale and divided into overlapping image patches. Note that the large-scale images are resized in such a way that their largest dimension does not exceed 400 pixels. Each of

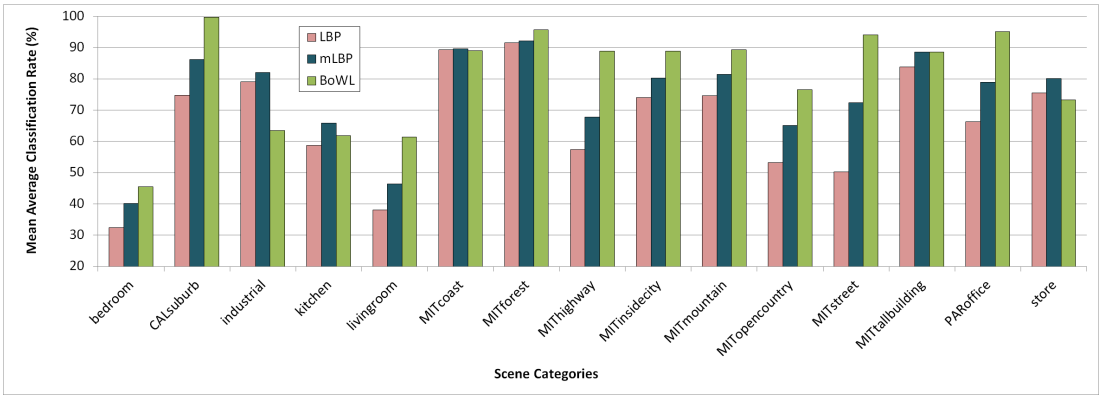


Figure 7.9 The comparative mean average classification performance of the LBP, mLBP and BoWL descriptors on the 15 categories of the Fifteen Scene Categories dataset.

these patches undergo the DCT-LBP process described in Section 7.1.2 to generate a bag of features. This bag of features is quantized using the k-means algorithm to form a visual vocabulary with 1000 words. Next each training and test image is represented as a pyramid histogram of these visual words. For evaluating the relative classification performances of the LBP, the mLBP and the BoWL descriptors, a Support Vector Machine (SVM) classifier with a Hellinger kernel (Vapnik 1995; Vedaldi and Fulkerson 2010; Arandjelović and Zisserman 2012) is used.

For the UIUC Sports Event dataset, 70 images are used from each class for training and 60 from each class for testing of the three descriptors. The results are obtained over five random splits of the data. As shown in Figure 7.8, the BoWL outperforms the LBP by a big margin of nearly 15%. BoWL also outperforms mLBP by over 13%. In fact, on this dataset the BoWL not only outperforms the LBP and mLBP, but also provides a decent classification performance on its own. Please note that the horizontal axis shows the different descriptors and the three datasets, and the vertical axis the mean average classification performance.

From both the MIT Scene dataset and the Fifteen Scene Categories dataset five random splits of 100 images per class are used for training, and the rest of the images are used for testing. Again, the BoWL produces decent classification performance on its own

Table 7.1 Comparison of the Classification Performance (%) of the Proposed Grayscale BoWL Descriptor with Other Popular Methods on the UIUC Sports Event, MIT Scene 8 and the 15 Scenes Datasets

Method		UIUC	Scene 8	15 Scenes
SIFT+GGM	(Li and Fei-Fei 2007)	73.4	-	-
OB	(Li et al. 2010)	76.3	-	-
KSPM	(Yang et al. 2009)	-	-	76.7
KC	(Van Gemert et al. 2010)	-	-	76.7
CA-TM	(Niu et al. 2012)	78.0	-	-
ScSPM	(Yang et al. 2009)	-	-	80.3
SIFT+SC	(Bo et al. 2011)	82.7	-	-
SE	(Oliva and Torralba 2001)	-	83.7	-
HMP	(Bo et al. 2011)	85.7	-	-
C4CC	(Bosch et al. 2006)	-	86.7	-
Grayscale BoWL		**87.7**	**91.6**	**80.7**

apart from beating the LBP and mLBP by a fair margin. Figure 7.8 displays these results on the MIT Scene dataset and Fifteen Scene Categories dataset. The highest classification rate for the MIT Scene dataset is as high as 91.6% for the BoWL descriptor which is a very good result for this dataset. The classification performance of BoWL beats that of LBP and mLBP both by a margin of over 17%. Please note that even without using color information, BoWL also beats the classification performance of the CLF and CGLF descriptors.

On the Fifteen Scene Categories dataset, the overall success rate for BoWL is 81.3% which is over 14% higher than LBP and over 6% higher than mLBP. This is also shown Figure 7.8. In Figure 7.9 the category wise classification rates of the grayscale LBP, the grayscale mLBP and the grayscale BoWL descriptors for all 15 categories of this dataset are shown. Here, the horizontal axis reveals the fifteen scene categories, and the vertical axis displays the mean average classification performance. The BoWL here is shown to better the LBP classification performance in 12 of the 15 scene categories and classify over 90% images correctly in four of the categories.

The classification performance of the proposed BoWL descriptor is also compared

with some popular image descriptors and classification techniques used by other researchers. The detailed comparison is shown in Table 7.1. It should be noted that the results of other researchers are reported directly from their published work.

7.3 Summary

In this chapter, a variation of the mLBP descriptor introduced in Chapter 3 is used with a DCT and bag-of-words based representation to form the novel Bag of Words-LBP (BoWL) image descriptor. This descriptor is used in conjunction with a spatial pyramid image representation and SVM classifier to test the classification performance. The experimental results on three popular datasets show that the BoWL descriptor significantly improves image classification performance over the LBP and the mLBP, and also yields classification performance at par with or better than several recent methods used by other researchers, such as the popular nonlinear Kernel Spatial Pyramid Matching (KSPM), SIFT Sparse-coded Spatial Pyramid Matching (ScSPM) and the Kernel Codebook (KC).

CHAPTER 8

CONCLUSIONS AND FUTURE WORK

This dissertation focuses on feature extraction from color and grayscale images by introducing several novel image descriptors based on texture, color, shape and local features. The main contributions of this dissertation are as listed below:

- A new color multi-mask Local Binary Patterns (mLBP) descriptor, which improves upon the traditional grayscale LBP, is introduced to represent texture information contained in an image for scene and texture image classification. Further, the multi-mask LBP descriptors from various color spaces and grayscale are combined to propose the new Color LBP Fusion (CLF) and the Color Grayscale LBP Fusion (CGLF) descriptors that perform well on texture and scene image datasets.

- An innovative HaarHOG descriptor is proposed for enhancing the HOG descriptor for encoding the shape and local features from an image. This was combined with an SVM classifier on four challenging datasets to show that the HaarHOG descriptor improves recognition performance over the HOG descriptor, as well as yields performance comparable to or better than several other popular descriptors on some datasets.

- Inspired by the LBP method, a novel Three Dimensional Local Binary Patterns (3D-LBP) descriptor is proposed, which uses the three color components to extract not only the texture but also the color feature from an image. Further, the 3D-LBP descriptor is combined with the HaarHOG descriptor to generate two new descriptors for color scene images — the 3DLH descriptor and the 3DLH-fusion descriptor. Results of the experiments using three challenging datasets show that the 3DLH-fusion descriptor improves recognition performance over several other popular descriptors. The fusion of multiple color 3DLH descriptors (3DLH-fusion) also shows an in-

crease in the classification performance, which suggests that various color 3DLH descriptors are not completely redundant for image classification.

- A novel H-descriptor is presented which integrates the 3D-LBP and the HOG of its Haar wavelet transform, to encode color, texture, shape, and local information from an image. Also the H-descriptor is comparatively assessed in seven different color spaces — the RGB, the HSV, the YCbCr, the oRGB, the $I_1I_2I_3$, the YIQ, and the DCS color spaces — for image classification performance. Further, the H-descriptor has been assessed in four randomly generated color spaces and grayscale to understand the role of specific color transformations and justify using color. Finally, a new H-fusion descriptor has been presented by fusing the PCA features of the H-descriptors in the seven color spaces. Experimental results using three datasets show that the proposed new H-fusion descriptor achieves better image classification performance than other popular descriptors, such as the Scale Invariant Feature Transform (SIFT), the Pyramid Histograms of visual Words (PHOW), the Pyramid Histograms of Oriented Gradients (PHOG), Spatial Envelope, Color SIFT four Concentric Circles (C4CC), Object Bank, the Hierarchical Matching Pursuit, as well as LBP, among others.

- Extending the mLBP method, a novel Bag of Words LBP (BoWL) descriptor is proposed, which combines DCT, mLBP and the bag-of-words representation techniques to drastically improve performance over the traditional LBP and mLBP. Results of the experiments using three challenging datasets show that the BoWL descriptor performs much better than the LBP and mLBP, and also performs at par with or better than several other popular descriptors.

Future work lies in the following directions:

- The proposed BoWL descriptor has only been tested on grayscale images. It can be extended to color images by splitting a color image into its color component images

and calculating the BoWL descriptor for each component. The BoWL descriptor needs to be tested on color images for a complete evaluation of its performance.

- All the descriptors proposed in this dissertation give equal weight to all regions of an image. Segmentation techniques exist that can detect objects within an image and give more weight to regions containing objects. It remains to be seen whether the descriptors presented here can further improve the classification performance in combination with such techniques.

REFERENCES

R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

S. Banerji, A. Verma, and C. Liu. Novel color LBP descriptors for scene and image texture classification. In *15th International Conference on Image Processing, Computer Vision, and Pattern Recognition*, pages 537–543, Las Vegas, Nevada, USA, July 18-21 2011.

G. Beylkin, R. Coifman, and V. Rokhlin. Fast wavelet transforms and numerical algorithms I. *Communications on Pure and Applied Mathematics*, 44(2):141–183, 1991.

L. Bo, X. Ren, and D. Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Neural Information Processing Systems*, pages 2115–2123, Granada, Spain, 2011.

A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *The European Conference on Computer Vision*, pages 517–530, Graz, Austria, 2006.

A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *The 11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, 2007a.

A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *International Conference on Image and Video Retrieval*, pages 401–408, Amsterdam, The Netherlands, July 9-11 2007b.

A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(4):712–727, 2008.

M. Bratkova, S. Boulos, and P. Shirley. oRGB: A practical opponent color space for computer graphics. *IEEE Computer Graphics and Applications*, 29(1):42–55, 2009.

G. Burghouts and J.-M. Geusebroek. Performance evaluation of local color invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009.

C.S. Burrus, R.A. Gopinath, and H. Guo. *Introduction to wavelets and wavelet transforms: A Primer*. Prentice-Hall, 1998.

B. Caputo, E. Hayman, and P. Mallikarjuna. Class-specific material categorisation. In *The Tenth IEEE International Conference on Computer Vision*, pages 1597–1604, Beijing, China, October 17-20, 2005.

W. Chen, M-J. Er, and S. Wu. Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 36(2):458–466, 2006.

M. Crosier and L.D. Griffin. Texture classification with a dictionary of basic image features. In *Computer Vision and Pattern Recognition*, pages 1–7, Anchorage, Alaska, June 23-28, 2008.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, pages 1–22, Prague, Czech Republic, 2004.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *The 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 886–893, San Diego, CA, USA, 2005.

R. Datta, D. Joshi, J. Li, and J. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):509–522, 2008.

L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *Conference on Computer Vision and Pattern Recognition*, pages 524–531, San Diego, CA, USA, 2005.

R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, Madison, Wisconsin, USA, June 16-22 2003.

M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, second edition, 1990.

T. Gevers, J. van de Weijer, and H. Stokman. Color feature detection: an overview. In Ratislav Lukac and Konstantinos N. Plataniotis, editors, *Color image processing: methods and applications*, chapter 9, pages 203–226. CRC Press, October 2006.

R.C. Gonzalez and R.E. Woods. *Digital Image Processing*. Pearson Prentice Hall, third edition, 2008.

G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007. URL `http://authors.library.caltech.edu/7694`.

J. Gu and C. Liu. Feature local binary patterns with application to eye detection. *Neurocomputing*, (0):–, 2013. ISSN 0925-2312. URL `http://www.sciencedirect.com/science/article/pii/S0925231213001513`.

Z. M. Hafed and M. D. Levine. Face recognition using the discrete cosine transform. *International Journal of Computer Vision*, 43(3):167–188, July 2001.

E. Hayman, B. Caputo, M. Fritz, and J-O. Eklundh. On the significance of real-world conditions for material classification. In *European Conference on Computer Vision*, pages 253–266, Prague, Czech Republic, May 11-14 2004.

Y. Jiang, C. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *The 6th ACM International Conference on Image and Video Retrieval*, pages 494–501, Amsterdam, The Netherlands, 2007.

S. Kondra and V. Torre. Texture classification using three circular filters. In *IEEE Indian Conference on Computer Vision, Graphics and Image Processing*, pages 429–434, Bhubaneshwar, India, December 16-19 2008.

S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2169–2178, New York, NY, USA, 2006.

L.-J. Li and L. Fei-Fei. What, where and who? classifying event by scene and object recognition. In *IEEE International Conference in Computer Vision*, pages 1–8, 2007.

L-J. Li, H. Su, E. P. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Neural Information Processing Systems*, pages 1378–1386, Vancouver, Canada, 2010.

C. Liu. A Bayesian discriminating features method for face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(6):725–740, 2003.

C. Liu. Enhanced independent component analysis and its application to content based face image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, 34(2):1117–1127, 2004.

C. Liu. Capitalize on dimensionality increasing techniques for improving face recognition grand challenge performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(5):725–737, 2006.

C. Liu. The Bayes decision rule induced similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(29):1116–1117, 2007.

C. Liu. Learning the uncorrelated, independent, and discriminating color spaces for face recognition. *IEEE Transactions on Information Forensics and Security*, 3(2):213–222, 2008.

C. Liu. Extracting discriminative color features for face recognition. *Pattern Recognition Letters*, 32(14):1796–1804, 2011.

C. Liu and V. Mago, editors. *Cross Disciplinary Biometric Systems*. Springer, 2012.

C. Liu and H. Wechsler. Robust coding schemes for indexing and retrieval from large face databases. *IEEE Transactions on Image Processing*, 9(1):132–137, 2000.

C. Liu and J. Yang. ICA color space for pattern recognition. *IEEE Transactions on Neural Networks*, 2(20):248–257, 2009.

Z. Liu and C. Liu. Fusion of the complementary discrete cosine features in the YIQ color space for face recognition. *Computer Vision and Image Understanding*, 111(3): 249–262, 2008.

D.G. Lowe. Object recognition from local scale-invariant features. In *The International Conference on Computer Vision*, pages 1150–1157, Corfu, Greece, 1999.

D.G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

O. Ludwig, D. Delgado, V. Goncalves, and U. Nunes. Trainable classifier-fusion schemes: An application to pedestrian detection. In *12th International IEEE Conference On Intelligent Transportation Systems*, volume 1, pages 432–437, St. Louis, USA, 2009.

S. Mallat. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

W. Niblack, R. Barber, and W. Equitz. The QBIC project: Querying images by content using color, texture and shape. In *SPIE Conference on Geometric Methods in Computer Vision II*, pages 173–187, San Diego, California, USA, 1993.

Z. Niu, G. Hua, X. Gao, and Q. Tian. Context aware topic model for scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2743–2750, Providence, RI, USA, June 16-21 2012.

E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. In *9th European Conference on Computer Vision*, pages 490–503, Graz, Austria, 2006.

Y. Ohta. *Knowledge-Based Interpretation of Outdoor Natural Color Scenes*. Pitman Publishing, London, 1985.

T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on Kullback discrimination of distributions. In *International Conference on Pattern Recognition*, pages 582–585, Jerusalem, Israel, 1994.

T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 29(1):51–59, 1996.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. In *The 1997 Conference on Computer Vision and Pattern Recognition*, pages 193–199, San Juan, Puerto Rico, 1997.

C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *The Sixth International Conference on Computer Vision*, pages 555–562, Bombay, India, 1998.

M. Pontil and A. Verri. Support vector machines for 3D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(6):637–646, 1998.

P. Porwik and A. Lisowska. The haar wavelet transform in digital image processing: Its status and achievements. *Machine graphics & vision*, 13(1):79–98, 2004.

J. Sanchez, F. Perronnin, and T. Campos. Modeling the spatial layout of images beyond spatial pyramids. *Pattern Recognition Letters*, 33(16):2216 – 2223, 2012.

B. Schiele and J. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

P. Shih and C. Liu. Comparative assessment of content-based face image retrieval in different color spaces. *International Journal of Pattern Recognition and Artificial Intelligence*, 19(7):1039–1048, 2005.

J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, Nice, France, 2003.

A.R. Smith. Color gamut transform pairs. *Computer Graphics*, 12(3):12–19, 1978.

H. Stokman and T. Gevers. Selection and fusion of color models for image feature detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3): 371–381, 2007.

M.J. Swain and D.H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

J.C. Van Gemert, C.J. Veenman, A.W.M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32 (7):1271–1283, 2010.

Y.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

A. Vedaldi and B. Fulkerson. Vlfeat — an open and portable library of computer vision algorithms. In *The 18th Annual ACM International Conference on Multimedia*, pages 1469–1472, Firenze, Italy, 2010.

A. Verma, S. Banerji, and C. Liu. A new color SIFT descriptor and methods for image category classification. In *International Congress on Computer Applications and Computational Science*, pages 819–822, Singapore, December 4-6 2010.

P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.

J. Yang, Y. Jiang, A.G. Hauptmann, and C. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Multimedia Information Retrieval*, pages 197–206, Augsburg, Bavaria, Germany, 2007.

J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1794–1801, Singapore, December 4-6 2009.

C. Zhang, J. Liu, J. Wang, Q. Tian, C. Xu, H. Lu, and S. Ma. Image classification using spatial pyramid coding and visual word reweighting. In *The 10th Asian conference on Computer vision*, pages 239–249, Queenstown, New Zealand, 2010.

J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007a.

J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, June 2007b.

L. Zhang, R. Chu, S. Xiang, S. Liao, and S. Z. Li. Face detection based on multi-block LBP representation. In *The 2007 International Conference on Advances in Biometrics*, pages 11–18, Seoul, Korea, 2007c.

W. Zhao, Y. Jiang, and C. Ngo. Keyframe retrieval by keypoints: Can point-to-point matching help. In *The Fifth International Conference on Image and Video Retrieval*, pages 72–81, Tempe, AZ, USA, 2006.

C. Zhu, C. Bichot, and L. Chen. Multi-scale color local binary patterns for visual object classes recognition. In *International Conference on Pattern Recognition*, pages 3065–3068, Istanbul, Turkey, August 23-26 2010.